Resumo da actividade da Linguateca de 1 de Janeiro de 2009 a 31 de Dezembro de 2009

Diana Santos, Luís Miguel Cabral e Luís Costa[1]

Dezembro de 2009

No ano de 2009 a Linguateca, como projecto temporário que foi, teve uma existência *sui generis*, visto que, se por um lado tentou manter a actividade que vinha a desenvolver desde 2000, nas suas três vertentes do modelo IRA (informação, recursos e avaliação), por outro tentou preparar-se para o futuro virando-se para áreas de aplicação mais relacionadas com projectos concretos, além de preparar um plano que espera aprovação.

Infelizmente os vários planos feitos ao longo do presente ano para a actividade da Linguateca foram sempre prejudicados pelo atraso considerável das fontes financiadoras em responder ou cumprir as promessas feitas. Assim, a 9 de Janeiro houve a primeira promessa de recomeçarmos "normalmente" em Fevereiro durante 6 meses, para haver uma nova Linguateca no segundo semestre deste ano. Contudo, o contrato referente a esse período provisório apenas foi assinado em Junho, e o período de seis meses foi convertido em um ano (que termina agora). Assim, as contratações, orçamentadas e planeadas para o primeiro semestre de 2009, tiveram como resultado que apenas durante os meses de Julho e Agosto tenhamos conseguido ter as quatro pessoas planeadas a trabalhar para a Linguateca (veja-se as datas de contratação dos relatórios em anexo). Depois os seus vínculos foram interrompidos, sempre na mira de vir a haver uma renovação, o que até agora não aconteceu. Também apenas a 1 de Outubro foi possível iniciar o contrato do Fernando Ribeiro na FCCN, que é o único elemento que tem futuro garantido na Linguateca por mais três anos à data da escrita do presente relatório.

Assim, este relatório, embora relate muito trabalho feito durante 2009, descreve uma situação que se deseja nunca venha a ser repetida, nomeadamente uma actividade quase frenética para obter resultados em pouco tempo, sempre à espera de uma resposta, mas de que não se conhece o futuro nem se podem fazer planos a médio prazo. Tal dá origem, aliás, a um desgaste e falta de confiança por parte da comunidade que servimos: muitos dos contactos e sugestões de colaboração foram ficando adiados à espera de se saber o futuro.

Em anexos, apresentamos o plano a ser executado pelo SINTEF, e os planos e respectivos relatórios dos quatro contratados a prazo pela Linguateca. Os três meses do novo colaborador na FCCN, o Fernando Ribeiro, visto terem sido praticamente usados na formação intensiva deste – tanto em relação às actividades correntes da Linguateca como ao RCAAP com que iniciámos colaboração – e na execução das várias tarefas de manutenção que os vários catálogos e a presença na rede da Linguateca exigem, não tornam necessário um relatório específico desse colaborador.

1. Avaliação

Uma das actividades mais importantes da Linguateca é a de proporcionar a possibilidade de sistemas de processamento do português poderem ser avaliados no âmbito de avaliações conjuntas, ou produzindo recursos para essa mesma avaliação.

1.1. GikiCLEF

Prosseguindo com a participação activa na co-organização de eventos internacionais de avaliação conjunta na área de processamento da língua, a Linguateca organizou a pista de avaliação conjunta intitulada "GikiCLEF: Cross-language Geographic Information Retrieval from Wikipedia", www.linguateca.pt/GikiCLEF/ no âmbito da pista maior de QA@CLEF do CLEF. O GikiCLEF veio dar continuidade ao GikiP, pista piloto que decorreu no ano de 2008.

[1] Os dois primeiros autores trabalharam praticamente a tempo inteiro para a Linguateca durante 2009, enquanto o terceiro esteve apenas afecto a tempo parcial (cerca de 65%) à Linguateca.

O GikiCLEF teve por objectivo avaliar sistemas que produziam respostas com ênfase especial sobre o domínio geográfico, sobre 10 línguas da Wikipédia (alemão, búlgaro, espanhol, inglês, italiano, neerlandês, norueguês nas suas duas variantes, português e romeno).

A organização desta pista envolveu a divulgação, definição da tarefa e das directivas de avaliação, preparação das colecções, criação dos tópicos e apresentação de resultados, assim como a escrita de vários artigos e apresentações (Santos & Cabral, 2009, 2010a,b, Santos et al., 2010c, Cardoso, 2009, 2010).

Com o propósito de distribuir os recursos e de proceder à avaliação dos resultados de forma eficiente e colaborativa foi desenvolvido o SIGA (Sistema de Gestão e Avaliação do GikiCLEF), um sistema colaborativo na rede que veio permitir alguns dos pontos acima, nomeadamente a criação de tópicos, avaliação e computação dos resultados bem como a sua divulgação.

O GikiCLEF culminou por fim com a disponibilização dos recursos usados, através do pacote GIRA, veja-se http://www.linguateca.pt/GIRA/, em que o código do sistema SIGA foi também disponibilizado.

1.2. ResPubliQA

ResPubliQA, uma nova pista de avaliação de sistemas de resposta automática a perguntas no domínio jurídico, foi organizada este ano pela primeira vez no âmbito do CLEF. A tarefa para os sistemas participantes consistiu em encontrar na colecção JRC-Acquis (contendo o texto de leis da União Europeia aplicáveis aos Estados Membros) passagens de texto onde se encontrassem as respostas a 500 questões criadas pela organização do evento.

A Linguateca contribuiu para este evento em duas vertentes:

- Na tradução das 500 perguntas de inglês para português de forma a possibilitar a participação de sistemas de resposta a perguntas em português. Isto envolveu também a participação na discussão sobre se algumas das perguntas criadas inicialmente pela organização seriam apropriadas.
- Na avaliação de dois conjuntos de resultados para servirem de referência.

1.3. Segundo HAREM

O livro do Segundo HAREM veio finalmente a lume em 2009, embora com data de 2008[2], e encontra-se publicado integralmente em http://www.linguateca.pt/LivroSegundoHAREM/. Contendo 450 páginas, 15 capítulos, nove apêndices, um glossário e um historial (Mota & Santos, 2008).

Além disso, foi apresentado um artigo sobre a pista piloto do ReRelEM no SEW 2009. (Freitas et al., 2009). Os recursos associados a esta tarefa foram além disso objecto de melhoria pela Cláudia Freitas.

Um artigo sobre o sistema SAHARA de avaliação do HAREM foi publicado no STIL (Gonçalo Oliveira & Cardoso, 2009). O pacote de recursos do Segundo HAREM, o LAMPADA assim como a própria avaliação conjunta foram descritos num resumo enviado para apreciação (Freitas et al., 2010).

1.4. Futuras avaliações conjuntas

No âmbito da preparação do futuro e da auscultação da comunidade na área, lançámos uma pergunta geral sobre quais as avaliações conjuntas que os investigadores e desenvolvedores de ferramentas de processamento para a língua portuguesa gostariam que tivessem lugar num futuro próximo, abrindo um debate no fórum da Linguateca e na lista consagrada à avaliação conjunta.

Embora algumas pessoas ligadas à Linguateca se pronunciassem por uma continuação do

[2] Tanto em relação ao pedido de ISBN como compromisso com os autores, o livro foi criado com a data de 2008, mas a sua existência em versão final só aconteceu no início de 2009.

HAREM/ReReLEM e GikiCLEF, ou uma mistura de ambos, a principal resposta que tivemos foi a de avaliar a extracção semi-automática de terminologia.

Contudo, enquanto não tivermos indicações sobre o futuro da Linguateca mais não podemos fazer do que assentar esta resposta.

2. Recursos

Outra das vertentes importantes do nosso trabalho é a manutenção dos recursos já disponibilizados pela Linguateca, a sua melhoria, a disponibilização de recursos criados por outrem, e a criação de novos, idealmente em conjunto com a comunidade.

2.1. AC/DC

Neste projecto foram criadas versões novas da anotação morfo-sintáctica (pelo PALAVRAS[3]) de praticamente todos os corpos. Os corpos foram também anotados com a informação semântica relativa a cor e roupa, tendo sido desenvolvidos vários programas dedicados a esse assunto. (Mota & Santos, 2009, 2010).

Os seguintes corpos foram aumentados ou melhorados no decorrer deste ano: O CONDIVport[4] assistiu à integração da parte dos textos sobre saúde, o corpo do Museu da Pessoa foi aumentado pela integração de entrevistas do Museu da Pessoa brasileiro, e o ECI-EBR foi dividido em excertos, cada excerto marcado com o seu género literário, e em alguns casos ainda suplementado com o tema do excerto.

Para além do habitual apoio aos utilizadores com dúvidas sobre a utilização do serviço, publicámos um artigo que foi apresentado como poster no STIL (Costa et al., 2009), assim como uma apresentação no RIVLS em Oslo (Santos, 2009g, 2010).

No âmbito da integração dos vários projectos da Linguateca, foi desenvolvido também o sistema VARRA em colaboração com o CISUC (Universidade de Coimbra) e a PUC-Rio (grupo CLIC) para validar relações semânticas com base em corpos, a ser usado por alunos de graduação em linguística na PUC-Rio (Freitas, 2009d).

Finalmente temos os corpos de três projectos no Brasil que nos foram oferecidos para disponibilizar também através do AC/DC: a FrameNet Brasil cedeu-nos um corpo de legendas[5]; o COMET os corpos técnicos, e a PUC-Rio o Corpobras. Contudo, não foi ainda possível integrá-los no AC/DC.

2.2. CorTrad

O CorTrad, um projecto conjunto COMET/NILC/Linguateca[6] de disponibilizar na rede vários corpos paralelos, cujo desenvolvimento começou em Maio de 2008, teve a sua primeira instalação no NILC ainda em 2008, mas foi já em 2009 que a primeira versão do seu sítio foi incluída no âmbito do

[3] O PALAVRAS foi desenvolvido por Eckhard Bick, e a sua referência chave é: Eckhard Bick, *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Aarhus, Denmark: Aarhus University Press, Novembro de 2000.

[4] O CONDIVport é um corpus de textos em português europeu e em português brasileiro, das décadas de 50, 70 e 90-2000, construído no âmbito do projecto *Convergência e Divergência no Léxico do Português*, financiado pela FCT (Refª POCTI/ LIN/48575/2002) e liderado por Augusto Soares da Silva, veja-se Augusto Soares da Silva, "Integrando a variação social e métodos quantitativos na investigação sobre linguagem e cognição: para uma sociolinguística cognitiva do português europeu e brasileiro", *Revista de Estudos Linguísticos* **16**, n. 1, Belo Horizonte, v. 16, jan./jun. 2008 pp. 49-81.

[5] A descrição completa reza: *O corpus LF (Legendas de Filmes), criado pelo Projeto FrameNet Brasil, http://www.framenetbr.ufjf.br, sediado na Universidade Federal de Juiz de Fora, http://www.ufjf.br, contém legendas de filmes em português do Brasil cedidas pelo portal OpenSubtitles.org.*

[6] Responsáveis por parte do COMET, do Departamento de Línguas Modernas da Universidade de São Paulo: Stella E.O. Tagnin e Elisa Duarte Teixeira; por parte do NILC, Sandra Aluísio.

projecto COMET na rede, já segundo o desenho de Patricia Tagnin, veja-se http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html

O CorTrad contém a esta data três subcorpos alinhados e marcados com subdivisões estruturais, mas em diferentes fases no que se refere à anotação sintáctica e semântica:

- Assim, o subcorpo jornalístico já se encontra anotado gramaticalmente em ambas as línguas, com o PALAVRAS e com o CLAWS, assim como anotado com o campo da cor e da roupa do lado português.
- O subcorpo técnico-científico da culinária também se encontra anotado na parte portuguesa, também com campos semânticos, e nas duas primeiras traduções para o inglês. A tradução final ainda não se encontra disponível
- O subcorpo literário, por seu lado, ainda está a ser revisto devido a problemas na codificação dos vários ficheiros que o compõem.

O CorTrad foi anunciado ao público em Novembro de 2009 e apresentado quer em Bergen (Tagnin et al., 2009) em Setembro como no Rio de Janeiro em Novembro (Teixeira et al, 2009). Está além disso em apreciação um outro artigo no LREC 2010 (Santos et al., 2010a), assim como Stella Tagnin tem feito várias apresentações nas suas aulas e palestras convidadas (Tagnin, 2009, 2010).

Convém referir que esta é a primeira vez que montamos um projecto noutra localização que não máquinas da Linguateca, o que se implica uma maior abrangência de diferentes necessidades também corresponde a um maior trabalho e profissionalização na instalação do dito. Existem assim dois lugares paralelos, o do desenvolvimento do CorTrad na Linguateca, e o público, no NILC, visível através do URL acima.

2.3. PAPEL

O PAPEL, Palavras Associadas Linguateca Porto-Editora, foi finalmente disponibilizado ao público seguindo o contrato que tínhamos com a Porto Editora, ou seja, seis meses após a entrega do mesmo a essa empresa. As versões 1.0 e 1.1 foram disponibilizadas num sítio criado para o efeito, http://www.linguateca.pt/PAPEL/, com muita documentação associada, assim como vários artigos foram publicados sobre uma primeira avaliação deste recurso (Gonçalo Oliveira et al, 2009a,b, 2010).

Embora, estritamente falando, o doutoramento do Hugo Gonçalo Oliveira não seja (pelo menos administrativamente falando) no âmbito da Linguateca, tem havido grande sinergia entre o CISUC e a Linguateca (e a Porto Editora), culminando aliás esta colaboração num artigo conjunto à APL de comparação de vários recursos (Santos et al., 2009b, 2010b).

Num ângulo académico, o Hugo fez a defesa pública da sua proposta de tese sobre ontologias para o português na Univ. de Coimbra tendo como primeira arguente Diana Santos.

Correntemente estamos a ultimar o sistema VARRA -- **V**alidação, **A**valiação e **R**evisão de **R**elações semânticas no **A**C/DC, usando as relações do PAPEL, numa colaboração entre a Linguateca, o CISUC e o Departamento de Letras da PUC-Rio, envolvendo o grupo de pesquisa em Linguística Computacional - CLIC - e alunos de graduação, como aliás já referido acima na secção sobre o AC/DC.

O Hugo também recentemente instalou um sistema interactivo de navegação no PAPEL (e outros recursos) denominado o Folheador.

2.4. Corpógrafo

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.

Durante este ano e devido ao facto de o Corpógrafo se encontrar no pólo do Porto e não ter havido ainda qualquer decisão sobre como os anteriores pólos pudessem continuar ou não, praticamente não houve desenvolvimentos em relação a este recurso. Assim, foram simplesmente realizadas algumas operações de manutenção relativas ao Corpógrafo, e pequenas melhorias: a inclusão

do processamento de mais um formato de documentos, o .DOCX, relativo ao Microsoft Office 2007 (e versões posteriores) e algumas novas funcionalidades ao nível da gestão de utilizadores, tais como a recuperação dos dados de acesso do utilizador, e a edição ou alterações dos dados do perfil do utilizador. Foi também criada uma lista de melhorias a serem executadas se e quando houver financiamento para tal.

2.5. Port4NooJ

O Port4NooJ (http://www.linguateca.pt/Repositorio/Port4NooJ/) é um conjunto de recursos linguísticos para o português desenvolvidos no âmbito do sistema NooJ, tendo em vista o processamento automático da nossa língua. Este conjunto de recursos, desenvolvido por Anabela Barreiro no âmbito do seu doutoramento e desde 2008 acessível do Repositório da Linguateca, está subjacente a vários outros recursos utilizados ou acessíveis da Linguateca, tais como o Corpógrafo e o ReEscreve.

No decorrer deste ano este recurso foi revisto e aumentado, tendo-se procedido à disponibilização das versões 2.0 e 2.1 deste recurso. Além disso, foram explicitadas várias relações semânticas – tais como sinonímia, hiponínima, resultado-de e acção-de – implicitamente incluídas no recurso, e tornadas disponíveis do sítio do Port4NooJ, http://www.linguateca.pt/Repositorio/Port4NooJ/relacoes_semanticas_explicitas.

2.6. ReEscreve

O ReEscreve, http://www.linguateca.pt/ReEscreve/, é um sistema de ajuda e sugestão à escrita em português que utiliza mecanismos de parafraseamento e edição de texto, desenvolvido por Anabela Barreiro no âmbito do seu doutoramento, em colaboração com a Linguateca.

Em 2009, o ReEscreve foi actualizado com a versão 2.0 do Port4NooJ assim como com uma nova versão do NooJ, tendo sofrido algumas alterações de acordo com alterações feitas nestes recursos. A sua interface foi também melhorada.

Houve também grande actividade na sua divulgação (Barreiro, 2009a,b,c, Barreiro & Cabral, 2009, 2010).

2.7. WPT 05

As colecções WPT, http://www.linguateca.pt/WPT/, consistem nas recolhas da web portuguesa no âmbito do *tumba!* criadas pelo grupo de investigação XLDB/LaSIGE da Faculdade de Ciências da Universidade de Lisboa no âmbito do pólo da Linguateca no XLDB (em funcionamento até dezembro de 2008) e disponibilizadas pela Linguateca.

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.

No início deste ano a Linguateca terminou a documentação deste recurso, que foi divulgado nas listas de discussão relacionadas com o processamento automático do português.

2.8. Outros recursos

Graças à tomada de contacto dos donos e criadores desses recursos com a Linguateca, colocámos no Repositório da Linguateca mais dois recursos importantes:

Erros ortográficos em português de Portugal e sua correcção, cedidos por José Carlos Medeiros: http://www.linguateca.pt/Repositorio/ErrosMedeiros/

Memórias de tradução português-inglês nas áreas de direito, indústria, história, gestão, geografia, engenharia, etc. http://www.linguateca.pt/Repositorio/MemoriasTraducao/ cedidas por uma tradutora com muitos anos de experiência (este processo ainda se encontra em curso à data da escrita do presente relatório).

3. Catálogo de publicações da Linguateca, SUPeRB e RCAAP

Desde 2007 que o catálogo de publicações da Linguateca se encontra implementado no SUPeRB (http://www.linguateca.pt/superb/), apresentando um leque de funcionalidades que veio auxiliar em muito a manutenção do serviço de catálogo de publicações da Linguateca, que, relembramos, tem por objectivo catalogar todas as publicações na área do processamento computacional do português.

No âmbito do desenvolvimento do SUPeRB, foi feita alguma revisão sobre a apresentação da interface em português e em inglês, assim como muitas melhorias em relação às funcionalidades oferecidas, tanto no que se refere à automatização da obtenção de listas de publicações como à usabilidade da adição de novas entradas.

Além disso, este ano pela primeira vez foi processada a exportação das actuais publicações da Linguateca de acesso aberto, 450 até à data, para o repositório RCAAP (http://www.RCAAP.pt), recorrendo ao Dublin Core, como primeira fase de colaboração com esta iniciativa.

3.1. Colaboração com o RCAAP

O portal RCAAP (Repositório Científico de Acesso Aberto de Portugal, http://www.rcaap.pt/) tem como objectivo a recolha, agregação e indexação dos conteúdos científicos em acesso aberto existentes nos repositórios institucionais das entidades nacionais de ensino superior, e outras organizações de investigação e desenvolvimento.

Dada a convergência de objectivos com as linhas mestras da Linguateca (total abertura e disponibilização livre de trabalhos) por um lado e por outro o trabalho feito no âmbito da Linguateca relacionado com a catalogação de publicações (catálogo de publicações sobre o processamento computacional do português e o sistema de gestão de referências bibliográficas SUPeRB) – de facto, pode dizer-se que, na nossa área limitada, a Linguateca foi uma precursora do RCAAP – iniciou-se este ano uma parceria entre a Linguateca e o projecto RCAAP.

Os primeiros resultados desta parceria foram a exportação do conteúdo aberto do catálogo de publicações sobre o processamento computacional do português para o RCAAP, como já mencionado, e o início pela Linguateca de trabalho preparatório com vista ao desenvolvimento de funcionalidades que possam dar um valor acrescentado ao portal do RCAAP.

Depois de várias reuniões e troca de mensagens com a equipa do RCAAP, ficou planeado para o início de 2010 uma colaboração formal entre estes dois projectos, em que a Linguateca desenvolveria funções avançadas de pesquisa em dados bibliográficos que poderiam ser aplicadas no RCAAP. Para tal o Fernando Ribeiro participou na conferência sobre acesso aberto que teve lugar na Universidade do Minho em Novembro.

3.2. Obtenção da lista de publicações da Linguateca

As publicações da Linguateca no período a que se refere o presente relatório encontram-se listadas no fim do mesmo (secção 5.0) por ordem estritamente cronológica, de acordo com os seguintes critérios:

1. Apresentamos as publicações que já vêm de trás (ver relatórios anteriores) cuja versão final apenas veio a lume este ano;
2. Apresentamos as publicações escritas e publicadas no ano transacto, incluindo as publicações que não foram criadas por membros da Linguateca mas que tiveram origem em eventos organizados pela Linguateca, marcadas com a indicação NO ÂMBITO;
3. Apresentamos as que se encontram ainda no prelo, e sobre as quais não temos portanto indicação de qual a referência final;
4. Finalmente também indicamos as publicações que enviámos para apreciação, mas que à data ainda não sabemos se virão ou não a ser publicadas nesses ou noutros canais.

Esta lista foi criada por Fernando Ribeiro com o auxílio do SUPeRB, e contém 78 publicações, das

quais 13 no âmbito da actividade da Linguateca (mas não produzidas por elementos da Linguateca). As primeiras 25 foram publicadas com a data de 2008; as últimas onze ainda não se encontram na sua forma definitiva.

4. Preparação do futuro e contactos vários

4.1. Envio de proposta da Linguateca a 3 anos

Conforme combinado e planeado, e depois de várias mensagens de correio electrónico a dar conta das ideias mestras, uma primeira proposta foi enviada a 5 de Agosto de 2009.

Nela se previa:

- Uma chamada para novos pólos da Linguateca
- Uma chamada para três iniciativas de vulto à roda das quais a nova Linguateca se estruturaria.

Infelizmente, até à data não tivemos qualquer resposta sobre a dita, o que prejudicou fortemente a definição da actividade.

4.2. Proposta de projecto KNOB

A proposta de projecto KNOB (KNowledge On Brazil) foi enviada à entidade que financia projectos de investigação na Noruega (Forskningsrådet). A ideia base subjacente a esta proposta é complementar e enriquecer trabalho já feito (ou a ser feito) pela Linguateca.

No âmbito deste projecto propusemos as seguintes três actividades principais:

- Organização de uma avaliação conjunta de sistemas de extracção de informação sobre a América Latina (na sequência lógica das iniciativas do GikiCLEF, GikiP e HAREM).
- Criação de corpos paralelos e comparáveis para o par português/norueguês utilizando e desenvolvendo a tecnologia desenvolvida nos projectos CorTrad, COMPARA e Squirrel.
- Desenvolvimento de um sistema de detecção de inconsistências sobre factos descritos em português, reutilizando, quando tal for apropriado, funcionalidades de outras ferramentas desenvolvidas no âmbito da Linguateca como o Esfinge e o REMBRANDT.

4.3. Participação noutras propostas e contactos vários

Devido à indefinição total em que o estatuto e a actividade da Linguateca se encontram e à falta de resposta atempada em relação a solicitações concretas, não pudemos infelizmente responder afirmativamente a muitas iniciativas para as quais fomos convidados ou quereríamos pertencer.

Dessa forma, tanto a comunicação com o projecto do arquivo da Web portuguesa como com a Biblioteca Nacional teve de esperar, assim como os contactos com Moçambique e com outros actores de processamento computacional de português em Portugal e no Brasil e a nível internacional foram sempre feitos indicando a nossa incerteza quanto ao futuro.

Relatamos contudo todas as colaborações ou contactos iniciados ou continuados neste ano, sob a chancela explícita de "preparação do futuro da Linguateca (ainda à espera de resposta das fontes financiadoras)", que não tenham ainda sido mencionados neste relatório:

- Contactos com os jornais *Notícias* de Moçambique e *Domingo* através de Armindo Ngunga e Christian-Emil Ore e com o projecto de colaboração entre a Univ. Eduardo Mondlane e a Universidade de Oslo que associa estes dois investigadores;
- Contactos com Knut Hofland e a Universidade de Bergen para usar ou melhorar o alinhador desenvolvido por este no âmbito do CorTrad;
- Contactos com Maria José Bocorny Finatto, e convite, no âmbito da Linguateca, (mas a expensas próprias) para visitar o pólo de Oslo em Agosto de 2010, por ocasião do curso

conjunto na ESSLLI 2010 a ser dado em Copenhaga nesse mês por Diana Santos e Maria José Bocorny Finatto;

- Contactos com Tony Berber Sardinha sobre a história dos corpos no Brasil;
- Contactos com Vládia Pinheiro no STIL 2009 e visita desta (a expensas próprias) ao pólo de Oslo da Linguateca em Outubro de 2009 para a apresentação de uma palestra sobre a sua tese de doutoramento em Fortaleza, em fase de conclusão, e para a eventual disponibilização da ConceptNet em português também através da Linguateca;
- Contactos com a FrameNet Brasil por ocasião da preparação da apresentação da Susana Afonso sobre a Floresta e sobre a FrameNet na Escola de Verão Belinda Maia em Junho-Julho de 2009;
- Contactos multilaterais com Max Silbertzein (envolvendo Anabela Barreiro, Belinda Maia, Cristina Mota e Diana Santos) na mira de uma colaboração futura entre o NooJ e a Linguateca;
- Contactos com Eckhard Bick para a preparação de uma nova versão do PALAVRAS a ser comparada pela Linguateca (à espera, infelizmente, desde Janeiro);
- Contactos com a empresa dinamarquesa/sueca Mikrov sobre corpos portugueses e erros ortográficos;
- Contactos com a Priberam e a Microsoft por ocasião de várias propostas de projectos destas duas empresas;
- Actividade regular de revisão de artigos para a *Linguamática* e outras revistas e conferências em português, e em inglês.

Além disso, refira-se a colaboração regular que tivemos ao longo do ano com a Porto Editora por causa do PAPEL, com a Universidade de Coimbra devido à escrita de muitos artigos conjuntos sobre o mesmo PAPEL e à relação privilegiada com o doutoramento do Hugo Gonçalo Oliveira, com o XLDB devido à co-orientação no âmbito do projecto GREASE e da Linguateca de Marcirio Chaves e de Nuno Cardoso. O primeiro defendeu a sua tese (Chaves, 2009) em Novembro, e o segundo a sua qualificação em Maio.

Também no âmbito do projecto AC/DC, contámos com a colaboração regular de Augusto Soares da Silva da Universidade Católica de Braga como consultor e fornecedor de dados (em relação ao CONDIVport). Conversamente, Diana Santos participou como consultora na proposta de projecto PERFIDE para alinhamento de corpos multilingues recentemente aprovada, graças ao convite de José João Dias da Silva e Alberto Simões.

Finalmente, houve uma colaboração significativa com a FLUP para a docência e concepção de vários cursos e palestras no âmbito da Escola de Verão Belinda Maia que teve lugar no Porto de 29 de Junho a 3 de Julho de 2009.

5. Anexo 1: descrição do trabalho a ser realizado pelo pólo de Oslo no SINTEF

Devido à precariedade de todo o processo, este ano assistiu a dois mini-contratos com o SINTEF, que incluímos aqui. Repare-se que os pontos 6) do primeiro contrato e 6) e 7) do segundo não puderam ser efectuados satisfatoriamente por falta de resposta das agências de financiamento.

5.1. Work description for the period from 2009-1-1 to 2009-6-30

Under the present contract SINTEF will continue to work on improving and maintaining a Portuguese resource (distributed) center for Portuguese language resources, Linguateca, having as main goal to ensure the existence and availability of resources such as corpora, electronic dictionaries and lexica, tools and an updated catalogue for Web materials, such as texts and relevant links for Portuguese.

In this period, and in addition to general maintenance and coordination of the work of the other nodes of Linguateca, the project is expected to develop the following activities:

1) GikiCLEF organization

2) significan improvements to the AC/DC project (both in terms of usability and by adding several relevant resources which have not te been made public)

3) CorTrad advancement and its dissemination

4) significant improvements to Corpógrafo

5) submission of papers about Linguateca to international journals

6) preparation of a longer-term project for 3 years

7) continuing supervision of PhD projects associated to RENOIR and PAPEL, corresponding to the doctoral theses of Nuno Cardoso and Hugo Oliveira

8) continuing Linguateca's infrastructure, performing monthly statistics, maintaining and improving the catalogue of publications, lists, and several Web services

9) collaboration with international entities or researchers with the purpose that they incorporate resources from Linguateca (Mike Scott - WordSmith Tools (Manchester), Armindo Ngunga (Mozambique), Knut Hofland - TCA2(Bergen), Stefan Evert - CQP (Osnabruck) and several Brazilian ones)

5.2. Work description for the period from 2009-7-1 to 2009-12-31

Under the present contract SINTEF will continue to work on improving and maintaining a Portuguese resource (distributed) center for Portuguese language resources, Linguateca, having as main goal to ensure the existence and availability of resources such as corpora, electronic dictionaries and lexica, tools and an updated catalogue for Web materials, such as texts and relevant links for Portuguese.

In this period, and in addition to general maintenance and coordination of the several researchers hired by Linguateca but not in Oslo, the project is expected to develop the following activities:

1) GikiCLEF organization: presentation of final results at CLEF workshop and paper writing

2) continuation of improvement in the AC/DC and CorTrad corpora projects

3) continuing Linguateca's infrastructure, performing monthly statistics, maintaining and improving the catalogue of publications, lists, and several Web services

4) submission of papers about Linguateca to international journals

5) continuation of PhD supervision associated to the subprojects RENOIR and PAPEL, corresponding to the doctoral theses of Nuno Cardoso and Hugo Oliveira

6) launching several new nodes in Linguateca, according to choice based on applications by the prospective partners

7) starting whatever new infrastructure projects that will have been approved in the proposal submitted as a result of the work in the previous period (described by amendment 7)

6. Anexo 2: planos e relatórios dos contratos a prazo

Por ordem de início da contratação, apresentamos os planos e respectivos relatórios de

- Cristina Mota, três meses a partir de 28 de Maio de 2009.
- Anabela Barreiro, três meses a partir de Junho de 2009.
- Rosário Silva, seis meses a 50% a partir de Junho de 2009
- Cláudia Freitas, seis meses a 30% a partir de Julho de 2009

Note-se, de qualquer maneira, que os planos de contratação eram relativamente vagos porque conhecíamos bem as investigadoras em questão e não sabíamos exactamente que actividades e que prazos teríamos dada a falta de resposta das entidades competentes. Além disso, pelo menos nos casos da Cristina Mota e da Cláudia Freitas algum do trabalho referido no plano e no relatório foi de facto executado antes da própria assinatura do contrato, ou seja, elas iniciaram o seu trabalho para a Linguateca antes de terem a sua situação resolvida em termos administrativos.

Os relatórios apresentados aqui foram ligeiramente abreviados em relação aos originais enviados à FCCN de forma a torná-los mais compatíveis com o estilo do presente texto, mas esta versão foi revista e aprovada pelas investigadoras em questão.

6.1. Cristina Mota: Plano

a) Colaborar na organização do GikiCLEF: análise dos resultados e programas de avaliação;
b) Construir recursos que estão em desenvolvimento, designadamente no AC/DC ou no CorTrad;
c) Redigir artigos e proceder à sua apresentação, em particular no Semantic Evaluation Workhsop, nos EUA a 4 de Junho;
d) Preparar o início da organização do Terceiro HAREM.

6.2. Cristina Mota: Relatório

Actividades relativas ao Segundo HAREM

- Conclusão da edição do livro Mota & Santos (2008) (revisão de conteúdo e formatação dos capítulos e apêndices, redacção do prefácio, revisão do glossário, criação de apêndices com resultados e descrição das colecções, revisão da bibliografia)
- Publicação do livro em formato digital na rede, e pedido de dois números ISBN (para edição digital e impressa) à APEL
- Contactos com autores e divulgação do livro em listas de distribuição nacional e internacional
- Preparação e apresentação do artigo sobre o ReRelEM (Freitas et al., 2009) na NAACL/SEW 2009, em Boulder, Colorado, EUA
- Actualização do sítio do Segundo HAREM

Actividades relativas ao GikiCLEF 2009

- Avaliação de respostas
- Revisão da documentação e testes de instalação ao sistema SIGA

Actividades relativas ao AC/DC

- Criação do programa `corte-e-costura` de aplicação de regras de anotação
- Testes ao programa corte-e-costura com criação de ficheiro auxiliar com regras-exemplo
- Documentação do programa (Mota & Santos, 2009)
- Escrita e envio de um resumo para apreciação sobre esse trabalho ao LREC 2010 (Mota & Santos, 2010)

Adicionalmente

- Criação de um pacote de latex com ficheiro de estilo da bibliografia em português para o STIL (a partir do pacote de edição da Linguateca)
- Participação, enquanto aluna, na *Escola de Verão Belinda Maia*, na FLUP.

6.3. Anabela Barreiro: Plano

1. melhoria significativa dos recursos por detrás do ReEscreve (Port4NooJ e gramáticas)
2. melhoria do texto (em português) da interface de vários recursos criados pela Linguateca: o próprio ReEscreve, o SUPeRB, e o Corpógrafo
3. participação na construção de recursos de avaliação da Linguateca, tal como o GikiCLEF e o HAREM
4. ensino sobre os recursos da Linguateca na Escola de Verão sobre "Aspectos do PLN em português" no Porto (29 de Julho a 3 de Julho de 2009)

6.4. Anabela Barreiro: Relatório

Preparação e docência dos seguintes cursos ou apresentações na *Escola de Verão Belinda Maia (Edv 2009)* (FLUP, Porto, Portugal, 29 de Junho a 3 de Julho de 2009):

- "Ajude-me a rever este texto! Editores estilísticos e ReEscreve" (Barreiro, 2009a).
- "Controlo da qualidade linguística e paráfrases: ajudantes da tradução automática" (Barreiro, 2009b)
- Anabela Barreiro & Cristina Mota. "Port4NooJ: até onde se pode ir?" (Barreiro & Mota, 2009)

Melhoria dos recursos linguísticos do Port4NooJ, recursos subjacentes ao ReEscreve (http://www.linguateca.pt/ReEscreve/):

- alargamento de propriedades morfossintácticas e semânticas nos dicionários e actualização e alargamento das gramáticas de parafraseamento.
- criação e/ou alargamento de ligações semânticas que resultam no parafraseamento de construções com verbos suporte de diversos tipos (construções predicativas nominais simples ou prepositivas, construções predicativas adjectivais/resultativas, variantes estilísticas das construções com verbos suporte elementares, etc.) e verbos plenos, entre locuções adverbiais e advérbios terminados em *-mente*, entre outras.

Melhoria do texto (em português) da página do Port4NooJ e da interface de alguns recursos criados na Linguateca: o próprio ReEscreve e o SUPeRB.
Disponibilização da versão 2.0 do Port4NooJ (com actualização da documentação) e das gramáticas de parafraseamento do ReEscreve (versão 3.0), e anúncio da nova versão do Port4NooJ nas listas.
Escrita ou participação na escrita de vários artigos (Barreiro & Cabral, 2009, 2010, Santos et al., 2010b), e apresentação de um deles no Canadá.
Participação na avaliação de respostas do GikiCLEF.

6.5. Rosário Silva: Plano

Revisão da anotação dos corpos da Linguateca
Participação no desenho de um sistema de apoio/ajuda a essa revisão

Colaboração na criação de recursos semânticos e de avaliação em curso na Linguateca, em particular avaliação do GikiCLEF e do HAREM
Escrita de artigos e apresentações quando tal se justificar

6.6. Rosário Silva: Relatório

Anotação semântica do AC/DC

- Anotação inicial das cores no CONDIVport, através de
    - Levantamento das palavras de cor na lista de adjectivos e verbos;
    - Identificação de mais cores com base nos campos semânticos «moda» e «roupa» e noutras pesquisas;
    - Actualização das listas de cor e dos grupos de cor já existentes e criação de novas listas mais específicas.
- Anotação da cor dos restantes corpos do projecto AC/DC:
    - Levantamento das palavras de cor na lista de adjectivos e verbos;
    - Identificação dos corpos onde constam as novas cores descobertas;
    - Actualização das listas de cor e dos grupos de cor já existentes;
    - Levantamento dos casos <mwe> de cor;
    - Redacção de algumas regras de ajuda à anotação automática da cor;
    - Elaboração da documentação referente à anotação da cor (Silva & Santos, 2009);
    - Identificação dos casos e redacção das regras para cor:equipa e cor:política;
    - Verificação da consistência das regras de ajuda à anotação nos corpos anotados com cor
- Sugestão de algumas relações semânticas relativamente aos lemas de cor para a comparação de ontologias lexicais
- Anotação semântica da roupa nos corpos do projecto AC/DC:
    - Análise dos grupos já existentes e proposta de novos grupos e de classes;
    - Levantamento de lemas dos grupos novos;
    - Levantamento de lemas com base na lista de lemas N do tema moda do Condiv;
    - Elaboração da documentação referente à anotação da roupa;
    - Actualização periódica de todos os ficheiros referentes à roupa.
- Melhoria do corpo ECI-EBR
    - Identificação dos géneros de texto existentes no ECI-EBR, respectiva marcação e posterior melhoria;
    - No caso de texto jornalístico, adição da secção do jornal; no caso de texto informativo, marcação do assunto;
    - Redacção de texto de documentação sobre essa marcação (Santos, Silva & Silva, 2009).

Avaliação de respostas no âmbito da avaliação conjunta GikiCLEF, em inglês e português.
Participação na escrita do artigo Santos et al. (2010b) e na apresentação no XXV Encontro Nacional da Associação Portuguesa de Linguística (Santos et al., 2009).

6.7. Cláudia Freitas: Plano

1. Preparação do Terceiro HAREM

a) Revisão das categorias e opções linguísticas do Segundo HAREM para uma eventual proposta para o Terceiro HAREM

b) Ampliação da marcação das relações entre EM, no sentido de vir a cobrir completamente a CD do HAREM

c) Eventual co-supervisão de uma bolsa de pesquisa na PUC de uma bolsista (nome retirado), orientada pela prof. Violeta Quental, relacionada com o HAREM e o ReRelEM, se essa bolsa for aprovada

2. Avaliação e uso dos corpos da Linguateca em ambiente de ensino

a) Em coordenação com as aulas de sintaxe na PUC-Rio, criação de exercícios e sua resolução com base na Floresta e no AC/DC e melhoria/compilação da lista de "Perguntas já respondidas sobre a utilização dos recursos da Linguateca"

3. Escrita de artigos e apresentações quando tal se justificar

6.8. Cláudia Freitas: Relatório

1. Preparação do Terceiro HAREM

Dado que não há de momento perspetivas de ser realizado um Terceiro HAREM, a atividade foi reduzida nesta área, com a concentração na parte das relações, e consequente viragem para a avaliação de relações semânticas também noutros recursos da Linguateca como o PAPEL.

Assim, e em relação aos recursos do ReRelEM, os 129 documentos da CD do Segundo HAREM foram anotados com as relações semânticas do ReRelEM, totalizando a anotação de relações semânticas entre mais de 2000 EM. Um dos objetivos da anotação era a validação das relações inicialmente propostas. Com a análise de mais textos, foi possível generalizar algumas relações, refinar e ainda criar outras que nos pareceram relevantes. Durante o processo de anotação, não foram poucas as dúvidas e discussões. Todo esse processo de anotação e discussão está documentado em http://sites.google.com/site/anotacaodorerelem/.

Com a aprovação da bolsa de pesquisa (nível de gradução), o trabalho da bolsista consistiu, em um primeiro momento, na familiarização com o HAREM e o ReRelEM (leituras e familiarizacao com o material das Colecções Douradas e com o programa Etiquet(h)arem). Em uma segunda fase do trabalho, ainda visando uma maior familiarização com as relações semânticas, foi feita uma comparação inicial entre as relações semânticas do ReRelEM e as relações semânticas presentes na WordNet.PT. Tal comparação teve como objectivo verificar em que medida as relações entre EM poderiam ser encampadas por relações semânticas mais gerais, tradicionalmente consideradas em recursos lexicais usados em tarefas que envolvem o processamento computacional da língua portuguesa.

2. Avaliação e uso dos corpos da Linguateca em ambiente de ensino

Quanto ao uso dos corpos da Linguateca em ambiente de ensino, foram trabalhadas questões envolvendo a coordenação de elementos no interior da frase – coordenação tanto no nível oracional quanto no nível do sintagma. Foi possível ainda apresentar o uso de corpos – e, especificamente, recursos e ferramentas que lidam com corpos anotados (como o AC/DC e a Floresta) – como auxiliares do professor de língua portuguesa na medida em que possibilitam a apresentação de fenómenos da língua em ambiente natural e facilitam a busca – por parte do professor – por esses mesmos fenômenos tendo em vista a elaboração de exercícios.

3. Escrita de artigos e apresentações

Participação na escrita dos artigos (Freitas et al., 2009, 2010, Santos et al., 2010b) e das apresentações (Freitas, 2009a,b,c, Santos et al., 2009).

4. Atividades relativas ao PAPEL

Em um primeiro momento, foram revisadas mais de 200 relações. Para dar continuidade ao processo de validação e torná-lo mais próximo à forma como processamos relações semânticas, foi criado o VARRA – Validação, Avaliação e Revisão de Relações semânticas no AC/DC, usando as relações do PAPEL. Nesse contexto, mencione-se a participação no desenvolvimento da interface e a escrita de um manual voltado aos utilizadores/validadores do VARRA (Freitas, 2009d).

7. Publicações da Linguateca no período referente a este relatório

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

2) [Amaral et al. 2008] 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, pp. 171-179. NO ÂMBITO

3) [Bruckschen et al. 2008] 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, pp. 247-260. NO ÂMBITO

4) [Cardoso 2008a] 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, pp. 195-211.

5) [Cardoso 2008b] Nuno Cardoso. "Apêndice G: 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, pp. 347-354.

6) [Carvalho & Freitas 2008] Paula Carvalho & Cláudia Freitas. "Apêndice D: Critérios de ALT". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008, pp. 319-321.

7) [Carvalho et al. 2008] 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, pp. 11-31.

8) [Carvalho & Gonçalo Oliveira 2008] Paula Carvalho & Hugo Gonçalo Oliveira. "Apêndice F: 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, pp. 339-346.

9) [Carvalho & Freitas 2008] Paula Carvalho & Cláudia Freitas. "Apêndice E: Exemplário do Segundo HAREM". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008, pp. 323-338.

10) [Chaves 2008] 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, pp. 231-245. NO ÂMBITO

11) [Craveiro et al. 2008] 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, pp. 159-170. NO ÂMBITO

12) [Ferreira et al. 2008] 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, pp. 213-229. NO ÂMBITO

13) [Freitas et al. 2008] 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, pp. 309-317.

14) [Freitas et al. 2008] 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, pp. 77-96.

15) [Gonçalo Oliveira et al. 2008] 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, pp. 97-129.

16) [Hagège et al. 2008a] 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, pp. 33-54.

17) [Hagège et al. 2008b] 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, pp. 289-308. NO ÂMBITO

18) [Martins 2008] 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, pp. 149-158. NO ÂMBITO

19) [Mota 2008] 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, pp. 181-193. NO ÂMBITO

20) [Mota et al. 2008a] 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, pp. 55-75.

21) [Mota et al. 2008b] Cristina Mota, Hugo Gonçalo Oliveira, Diana Santos, Paula Carvalho & Cláudia Freitas. "Apêndice I: Resumo de resultados do Segundo HAREM". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008, pp. 379-403.

22) [Mota et al. 2008c] Cristina Mota, Diana Santos, Paula Carvalho, Cláudia Freitas & Hugo Gonçalo Oliveira. "Apêndice H: Apresentação detalhada das colecções do Segundo HAREM". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008, pp. 355-377.

23) [Santos 2008] 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, pp. 1-7.

24) [Santos et al. 2008] 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, pp. 131-146.

25) [Santos et al. 2008] 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, pp. 277-286.

26) [Freitas 2009a] Cláudia Freitas. "Linguística Computacional, corpus, pesquisa e ensino da língua portuguesa". (PUC-Rio, Brasil, 16 de Abril de 2009).

27) [Santos 2009a] Diana Santos. "Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva". *Linguamática* **1.1** (2009), pp. 25-59.

28) [Freitas et al. 2009a] Cláudia Freitas, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira & Paula Carvalho. "Detection of relations between named entities: report of a shared task". In *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, SEW-2009* (Boulder, Colorado, USA, June 4, 2009), pp. 129-137.

29) [Afonso & Santos 2009] Susana Afonso & Diana Santos. "Florestas (treebanks) - apresentação geral, com especial relevo para a Floresta Sintá(c)tica". *Escola de Verão Belinda Maia (Edv 2009)* (FLUP, Porto, Portugal, 29 de Junho - 3 de Julho 2009).

30) [Barreiro 2009a] Anabela Barreiro. "Ajude-me a rever este texto! - Editores estilísticos e ReEscreve". *Escola de Verão Belinda Maia (Edv 2009)* (FLUP, Porto, Portugal, 29 de Junho - 3 de Julho 2009).

31) [Barreiro & Mota 2009] Anabela Barreiro & Cristina Mota. "Port4NooJ: até onde se pode ir?". *Escola de Verão Belinda Maia (Edv 2009)* (FLUP, Porto, Portugal, 29 de Junho - 3 de Julho 2009).

32) [Barreiro 2009b] Anabela Barreiro. "Controlo da qualidade linguística e paráfrases: ajudantes da tradução automática". *Escola de Verão Belinda Maia (Edv 2009)* (FLUP, Porto, Portugal, 29 de Junho - 3 de Julho 2009).

33) [Santos 2009b] Diana Santos. "Para inglês ver: Questões fundamentais e uma abordagem nova". *Escola de Verão Belinda Maia (Edv 2009)* (FLUP, Porto, Portugal, 29 de Junho - 3 de Julho 2009).

34) [Santos 2009c] Diana Santos. "POS tagging: clarificação histórico-terminológica". *Escola de Verão Belinda Maia (Edv 2009)* (FLUP, Porto, Portugal, 29 de Junho - 3 de Julho 2009).

35) [Santos & Cardoso 2009] Diana Santos & Nuno Cardoso. "REMando para o futuro: reconhecimento de entidades mencionadas e não só". *Escola de Verão Belinda Maia (Edv 2009)* (FLUP, Porto, Portugal, 29 de Junho - 3 de Julho 2009).

36) [Santos 2009d] Diana Santos. "Pesquisando corpos: Foi você que pediu um Corpo Ferreira?". *Escola de Verão Belinda Maia (Edv 2009)* (FLUP, Porto, Portugal, 29 de Junho - 3 de Julho 2009).

37) [Santos 2009e] Diana Santos. "Relação entre informática e linguística, ou. entre informáticos e linguistas". *Escola de Verão Belinda Maia (Edv 2009)* (FLUP, Porto, Portugal, 29 de Junho - 3 de Julho 2009).

38) [Santos 2009f] Diana Santos. "Linguateca: objectivos e resultados". *Encontro Ciência 2009* (Fundação Calouste Gulbenkian, Lisboa, Portugal, 29 e 30 de Julho de 2009).

39) [Barreiro & Cabral 2009] Anabela Barreiro & Luís Miguel Cabral. "ReEscreve: a translator-friendly multi-purpose paraphrasing software tool". In Marie-Josée Goulet, Christiane Melançon, Alain Désilets & Elliott Macklovitch (eds.), *Proceedings of the Workshop Beyond Translation Memories: New Tools for Translators, The Twelfth Machine Translation Summit* (Château Laurier, Ottawa, Ontario, Canada, 29 August 2009), pp. 1-8.

40) [Costa et al. 2009] Luís Costa, Diana Santos & Paulo Alexandre Rocha. "Estudando o português tal como é usado: o serviço AC/DC". In *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)* (São Carlos, Brasil, 8-11 de Setembro de 2009).

41) [Gonçalo Oliveira et al. 2009a] Hugo Gonçalo Oliveira, Diana Santos & Paulo Gomes. "Avaliação da extracção de relações semânticas entre palavras portuguesas a partir de um dicionário". In *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)* (São Carlos, Brasil, 8-11 de Setembro de 2009).

42) [Gonçalo Oliveira & Cardoso 2009] Hugo Gonçalo Oliveira & Nuno Cardoso. "SAHARA: an online service for HAREM Named Entity Recognition Evaluation". In *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)* (São Carlos, Brasil, 8-11 de Setembro de 2009).

43) [Santos 2009g] Diana Santos. "Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties". *Workshop on research infrastructure for linguistic variation* (University of Oslo, 17-18 September 2009).

44) [Cardoso 2009] Nuno Cardoso. "GikiCLEF topics and Wikipedia articles: did it blend?". *CLEF2009* (Corfu, Greece, 30 September - 2 October).

45) [Dornescu 2009] Iustin Dornescu. "EQUAL - Encyclopaedic QA for Lists". *GikiCLEF overview session at CLEF workshop (GikiCLEF)* (Corfu, Greece, 30 September - 2 October). NO ÂMBITO

46) [Hartrumpf & Leveling 2009] Sven Hartrumpf & Johannes Leveling. "GIRSA-WP at GikiCLEF: Integration of Structured Information and Decomposition of Questions". *GikiCLEF overview session at CLEF workshop (GikiCLEF)* (Corfu, Greece, 30 September - 2 October). NO ÂMBITO

47) [Larson 2009] Ray R. Larson. "Interactive Probabilistic Search for GikiCLEF". *GikiCLEF overview session at CLEF workshop (GikiCLEF)* (Corfu, Greece, 30 September - 2 October). NO ÂMBITO

48) [Santos & Cabral 2009] Diana Santos & Luís Miguel Cabral. "GikiCLEF: Crosscultural issues in an international setting: asking non-English-centered questions to Wikipedia". In Francesca Borri, Alessandro Nardi & Carol Peters (eds.), *Cross Language Evaluation Forum: Working notes for CLEF 2009* (Corfu, Greece, 30 September - 2 October).

49) [Tagnin et al. 2009] Stella O. E. Tagnin, Elisa Duarte Teixeira & Diana Santos. "CorTrad: a multiversion translation corpus for the Portuguese-English pair". *Arena Romanistica* **4** (2009), pp. 316-323. [The 28th International Conference on lexis and grammar, Bergen, Norway, 30 September - 3 October 2009]

50) [Chaves 2009] Marcirio Silveira Chaves. Uma Metodologia para Construção de Geo-Ontologias. Doutoramento. Faculdade de Ciências, Universidade de Lisboa. Setembro de 2009.

51) [Costa 2009] 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, pp. 361-368.

52) [Forner et al. 2009] 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, pp. 262-295.

53) [Mandl et al. 2009] 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, pp. 808-821.

54) [Mota & Santos 2009] Cristina Mota & Diana Santos. "Corte e costura no AC/DC: auxiliando a melhoria da anotação nos corpos". Setembro de 2009.

55) [Santos et al. 2009a] 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, pp. 894-905.

56) [Freitas 2009b] Cláudia Freitas. "Floresta Sintática". (PUC-Rio, Brasil, 6 de Outubro de 2009).

57) [Gonçalo Oliveira et al. 2009b] Hugo Gonçalo Oliveira, Diana Santos & Paulo Gomes. "Relations extracted from a Portuguese dictionary: results and first evaluation". In Luís Seabra Lopes, Nuno Lau, Pedro Mariano & Luís M. Rocha (eds.), *New Trends in Artificial Intelligence, Local Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA 2009)* (Aveiro, Portugal, October 12-15, 2009), pp. 541-552.

58) [Santos et al. 2009b] Diana Santos, Anabela Barreiro, Luís Costa, Cláudia Freitas, Paulo Gomes, Hugo Gonçalo Oliveira, José Carlos Medeiros & Rosário Silva. "O papel das relações semânticas em português: Comparando o TeP, o MWN.PT e o PAPEL". *XXV Encontro*

Nacional da Associação Portuguesa de Linguística (Lisboa, Portugal, 22-24 de Outubro de 2009).

59) [Freitas 2009c] Cláudia Freitas. "Apresentação da Linguateca com ênfase nos recursos". (PUC-Rio, Brasil, 4 de Novembro de 2009).

60) [Tagnin 2009] Stella E. O. Tagnin. "Projeto COMET - Avanços e Carências". *1º. Workshop de Linguística Computacional, FFLCH / USP* (São Paulo, Brasil, 17 Novembro 2009). NO ÂMBITO

61) [Teixeira et al. 2009] Elisa D. Teixeira, Diana Santos & Stella E. O. Tagnin. "CorTrad: um novo corpus paralelo multiversão para o par de línguas português-inglês". In *VIII Encontro de Linguistica de Corpus (ELC 2009)* (Rio de Janeiro, Brasil, 13-14 de Novembro de 2009).

62) [Freitas 2009d] Cláudia Freitas. "Instruções de validação de relações semânticas entre palavras", 18 de Dezembro de 2009.

63) [Silva & Santos 2009] Rosário Silva & Diana Santos. "Arco-íris: notas sobre a anotação do campo semântico da cor em português". Primeira edição: 25 de Junho de 2009. Em edição permanente.

64) [Santos, Silva & Silva 2009] Diana Santos, Rosário Silva & Augusto Soares da Silva. "Guarda-fatos: notas sobre a anotação do campo semântico do vestuário em português". Primeira edição: 26 de Outubro de 2009. Em edição permanente.

65) [Santos & Cabral 2010a] Diana Santos & Luís Miguel Cabral. "GikiCLEF : Expectations and lessons learned". In Carol Peters et al (ed.), *CLEF post-workshop proceedings* 2010, Springer. No prelo.

66) [Cardoso 2010] Nuno Cardoso. "GikiCLEF topics and Wikipedia articles: Did they blend?". In Carol Peters et al (ed.), *CLEF post-workshop proceedings* 2010, Springer. No prelo.

67) [Teixeira et al. 2010] Elisa D. Teixeira, Diana Santos & Stella E. O. Tagnin. "CorTrad: um novo corpus paralelo multiversão para o par de línguas português-inglês". In Tania Shepherd, Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Linguística de Corpus: Sínteses e Avanços. Anais do VIII Encontro de Linguística de Corpus, realizado na UERJ, 13 a 14 de novembro de 2009.* (Rio de Janeiro, RJ, 2010). No prelo.

68) [Freitas et al. 2010] Cláudia Freitas, Paula Carvalho, Hugo Gonçalo Oliveira, Cristina Mota & Diana Santos. "Second HAREM: advancing the state of the art of named entity recognition in Portuguese". In *The seventh international conference on Language Resources and Evaluation (LREC 2010)* (Malta, 10-21 Maio 2009). Enviado para apreciação.

69) [Mota & Santos 2010] Cristina Mota & Diana Santos. "Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora". In *The seventh international conference on Language Resources and Evaluation (LREC 2010)* (Malta, 10-21 Maio 2009). Enviado para apreciação.

70) [Santos & Cabral 2010b] Diana Santos & Luís Miguel Cabral. "GikiCLEF: Crosscultural issues in multilingual information access". In *The seventh international conference on Language Resources and Evaluation (LREC 2010)* (Malta, 10-21 Maio 2009). Enviado para apreciação.

71) [Santos et al. 2010a] Diana Santos, Stella E. O. Tagnin & Elisa Duarte Teixeira. "Colours, clothing and food in CorTrad: why corpus-based translation studies are revealing". In *The seventh international conference on Language Resources and Evaluation (LREC 2010)* (Malta, 10-21 Maio 2009). Enviado para apreciação.

72) [Barreiro & Cabral 2010a] Anabela Barreiro & Luís Miguel Cabral. "ReWriter: a New Language Composition Tool for English". In *11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2010)* (Iasi, Roménia, 21-27 de Março 2010). Enviado para apreciação

73) [Santos et al. 2010b] Diana Santos, Anabela Barreiro, Cláudia Freitas, Hugo Gonçalo Oliveira, José CarlosMedeiros, Luís Costa, Paulo Gomes & Rosário Silva. "Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL". In *Textos seleccionados apresentados ao XXV Encontro Nacional da Associação Portuguesa de Linguística*. APL, 2010. Enviado para apreciação

74) [Gonçalo Oliveira et al. 2010] Hugo Gonçalo Oliveira, Diana Santos & Paulo Gomes. "Extracção de relações semânticas entre palavras a partir de um dicionário: primeira avaliação". *Linguamática* (2010). Enviado para apreciação

75) [Santos 2010] Diana Santos. "Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties". *Oslo Studies in Language* (2010). Enviado para apreciação.

76) [Santos et al. 2010c] Diana Santos, Nuno Cardoso & Luís Miguel Cabral. "How geographical was GikiCLEF? A GIR-critical review". *GIR 2010* 2010. Enviado para apreciação

77) [Tagnin 2010] Stella E. O. Tagnin. "CorTrad - um corpus de traduções inglês-português online com mil e uma possibilidades de pesquisa!". *10o. Encontro de Férias da SBS* (14 a 16 de Janeiro de 2010). Em preparação. NO ÂMBITO

78) [Santos & Finatto 2010] Diana Santos & Maria José Bocorny Finatto. "Words and their secrets", ESSLLI 2020. Em preparação.

Índice