Resumo da actividade da Linguateca de 1 de Janeiro de 2010 a 31 de Dezembro de 2010

Diana Santos, Fernando Ribeiro, Cristina Mota, Cláudia Freitas e Rosário Silva[1]

Dezembro de 2010

O ano de 2010 foi traumatizante para a Linguateca, visto que condições fora do controle das várias partes levaram à impossibilidade de manter um contrato com o SINTEF, que foi transformado num subsídio de 3 meses junto com uma contratação direta da líder do projeto pela FCCN, numa afetação de 20%.

À data da escrita deste relatório não é ainda também conhecido o futuro deste projeto. Contudo, relatamos o trabalho feito na esperança de que isso leve a uma continuação do projeto mesmo que em moldes diferentes.

A atividade principal da Linguateca no presente ano centrou-se nas seguintes atividades

1. Desenvolvimento de ferramentas para estudar e eventualmente melhorar o conteúdo do RCAAP
2. Melhoramento do AC/DC e dos subprojetos a ele associados, tal como o VARRA e o corte-e-costura
3. Retoma da Floresta, em especial a Amazónia, e do CorTrad como recursos cujo desenvolvimento é necessário

Além disso mantivemos na medida das nossas possibilidades os serviços de repositório, catálogo e fórum da Linguateca, com melhorias a nível das estatísticas e do SUPeRB, assim como também demos apoio a mais uma avaliação conjunta de resposta automática a perguntas, a ResPubliQA 2010.

Apresentamos separadamente os relatórios individuais das três contratadas, visto que o seu contrato o estipula, assim como o relatório relativo ao subsídio ao SINTEF que foi também necessário produzir em Maio. A lista de publicações e apresentações durante o ano de 2010 é também apresentada.

1. 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 foram feitas várias melhorias em relação às funcionalidades oferecidas, tanto no que se refere à automatização da limpeza dos dados (nomes de autores e editores) como à usabilidade da adição de novas entradas quando se referem a "eventos".

1.1. 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. Como já frisámos no relatório anterior, existe uma notável convergência de objectivos entre a Linguateca e o RCAAP: veja-se as linhas mestras da Linguateca (total abertura e disponibilização livre de trabalhos) por um lado e, por outro, o trabalho de

[1] A afetação dos diferentes autores à Linguateca foi a seguinte: a primeira trabalhou a 100% até Agosto, e 20% a partir daí; o segundo esteve totalmente afeto, e a terceira, quarta e quinta autoras tiveram contratos que corresponderam a cerca de 80%, 33% e 25% respetivamente.

catalogação de publicações na área do processamento do português e o sistema de gestão de referências bibliográficas SUPeRB, que qualifica a Linguateca como uma espécie de precursora do RCAAP.

Dado isso, uma primeira colaboração ocorreu este ano, tendo sido desenvolvidos pela Linguateca vários estudos do material presente no RCAAP e sugestão de como melhorar este serviço com recursos e ferramentas de processamento do português.

Esta primeira fase consistiu basicamente na análise dos diários (logs) e dos metadados do RCCAP, em especial os nomes dos autores.

Em relação aos metadados, criou-se uma lista de nomes de autores e fez-se a sua uniformização. Este passo consistiu em escrever os nomes dos autores de uma forma igual para todas as entradas:

- removendo números nos nomes,
- reescrevendo o nome na forma nome de baptismo seguido dos diversos apelidos,
- separando as iniciais do nome e colocando um ponto numa inicial que não tenha ponto.

Como exemplo: Santos, Diana passou a Diana Santos e Cortez, Paulo, 1971- passará a Paulo Cortez.

Seguidamente fez-se uma ambiguação dos nomes, que consiste em representar o nome de várias formas. Os utilizadores ao introduzirem numa pesquisa um nome nem sempre o fazem introduzindo o nome completo, pelo que este passo serve para possíveis identificações na representação de um nome e obter um maior número de possíveis combinações desse nome. Como exemplo, a ambiguação do nome Diana Maria Santos, dará como resultado da ambiguação:

- D. Maria Santos,
- D. M. Santos,
- Diana M. Santos,
- Diana Maria Santos.

Seguidamente fizeram-se agrupamentos (clusters) para os nomes: um agrupamento em que se definiu cada grupo com a primeira letra do nome e o apelido, por exemplo*, J-Almeida*, para os nomes que comecem por J e terminem em Almeida e um outro agrupamento de grupos máximos que consiste na maior sequência de nomes, por exemplo, *Jorge Fernando Marques de Almeida* e *José Carlos Martins de Almeida*. Estes nomes fazem parte do mesmo grupo J-Alemida, mas pertencem a grupos diferentes para os grupos máximos, um grupo será *Jorge Fernando Marques de Almeida* e o outro *José Carlos Martins de Almeida*. Estes grupos vão conter elementos com nomes mais pequenos, ou seja, *J. Almeida* pertencerá aos dois grupos mas *J. F. Almeida* só pertencerá ao grupo *Jorge Fernando Marques de Almeida*.

Seguidamente fez-se uma análise aos dados obtidos dos diários e dos metadados com os dados uniformizados fazendo a associação de cada um dos nomes com os grupos criados. Finalmente usámos a ferramenta *JSpell* para correcção de erros e para dar sugestões de nomes para uma determinada entrada.[2] Este trabalho pode ser consultado em Santos & Ribeiro (2010), referente ao estado do RCAAP em Junho. Estamos de momento a desenvolver um serviço da rede que permite consultar as estatísticas atualizadas em relação ao RCAAP, com uma atualização mensal, além de permitir invocar os vários módulos de análise. Uma versão preliminar pode ser vista em http://www.linguateca.pt/colabRCAAP/.

Também no âmbito da cooperação Linguateca-RCAAP, o Fernando Ribeiro esteve presente nas reuniões de trabalho do RCCAP ao longo do ano, nomeadamente em Leiria (Março), Porto (Junho), Santarém (Outubro) e Braga (Novembro).

[2] A versão do *JSpell* usada é uma modificação da versão que está publicamente disponível para permitir definir um dicionário, não de termos isolados, mas de nomes de autores completos. As alterações foram feitas por Alberto Simões, mas esta versão ainda se encontra numa fase inicial de desenvolvimento.

1.2. 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 6) 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 transato;
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 28 publicações; das quais as últimas cinco ainda não se encontram na sua forma definitiva.

2. Avanços no projeto AC/DC

Podemos referir três vertentes:

- Atualização dos corpos e melhorias pontuais: alguma limpeza de problemas, assim como a correção do caso dos duplos clíticos, e aumento de atributos estruturais, quer em relação à variante, quer em relação a temas e ou fontes dos diversos materiais. Além disso, o corpo Floresta foi incluído no AC/DC.
- Anotação semântica: todos os corpos foram anotados com cor e roupa, e grande parte do material sobretudo em relação à cor foi revisto, assim como o pacote corte-e-costura foi disponibilizado
- VARRA: melhoria do sistema e da sua documentação e início do seu uso com alunos da PUC-Rio

3. Continuação do desenvolvimento de recursos

Muito resumidamente, indicamos aqui a continuação do desenvolvimento de variados recursos.

- A atividade da disponibilização das memórias de tradução foi completada.
- Nova versão do PAPEL assim como a lista das relações presentes no PAPEL que existiam nos maiores corpos do AC/DC e no Google foram adicionadas ao sítio do PAPEL.
- Amazónia parcialmente revista, e melhoria significativa do sítio da Floresta.
- Lista de erros ortográficos e sua correção também disponibilizada.
- Avanços no projeto CorTrad: início do desenvolvimento de mais dois subcorpos, um de tratados internacionais, outro de resumos de teses, assim como início da revisão do alinhamento do subcorpo jornalístico.

4. Outras atividades

Continuámos a manutenção do sítio da Linguateca, como adições ao catálogo, manutenção do fórum, atualização mensal das estatísticas de acesso e sua melhoria, assim como foi executada a migração do recurso da Geo-Net-Pt 02 do XLDB para a Linguateca. Tivemos além disso de proceder à reposição das páginas antigas dos pólos por causa de uma auditoria que nos tornou conscientes da necessidade de manter as páginas antigas por um período de cinco anos.

O SUPeRB continuou a ser desenvolvido e melhorado: adicionaram-se novas funcionalidades e alguns problemas foram resolvidos. Em particular, foram levados a cabo ou implementados os pontos

seguintes:

- Adição de ajuda para inserção de publicações,
- Inserção/edição das publicações no formato BibTeX
- Antevisão das referências antes de as guardar
- Limpeza geral dos nomes dos autores

A orientação do doutoramento do Nuno Cardoso continuou ao longo de todo o ano, e consequente colaboração com o XLDB para esse efeito, assim como o contato estreito com a Universidade de Coimbra por causa do PAPEL.

Contatos com o projeto PERFIDE também levaram à participação no encontro PERFIDE.

De um ponto de vista operacional, convem referir que este ano foi adquirido e instalado um novo servidor para a Linguateca, que entrou em produção em Março. Foram feitas as migrações da maior parte dos serviços do servidor antigo para o novo, de forma a que a transição fosse transparente aos utilizadores da Linguateca.

Foram também necessárias algumas intervenções no servidor da Linguateca alojado na REFER (servidor antigo), nomeadamente para a reinicialização da máquina em algumas ocasiões, o que levou a que em Novembro este servidor da Linguateca fosse realojado nas instalações da FCCN para evitar estas deslocações.

5. Anexo 1: relatório do trabalho realizado pelo pólo de Oslo no SINTEF

O trabalho contemplado no protocolo referente ao subsídio, assinado entre o SINTEF e a FCCN, continha os seguintes pontos:

- Melhoria considerável do projecto AC/DC
- Primeira versão estável do projecto CorTrad
- Continuação do trabalho associado aos projectos de doutoramento de Nuno Cardoso e de Hugo Oliveira (PAPEL)
- Envio de dois artigos sobre o trabalho da Linguateca a revistas internacionais

Podemos considerar que o trabalho foi todo efectuado, e foi descrito em mais pormenor no relatório em inglês enviado pelo SINTEF a 31 de Maio de 2010.

Não houve desvios excepto os procedentes das datas de publicação de artigos científicos, que são sempre mais morosas do que inicialmente esperado. Assim, o artigo Diana Santos. “Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties”. *OSLa*, ainda não foi publicado, embora já aceite por essa revista.

6. Anexo 2: relatórios dos contratos a prazo

- Por ordem de início da contratação, apresentamos os relatórios deCristina Mota, nove meses a 80%
- Rosário Silva, doze meses a 25%
- Cláudia Freitas, doze meses a 30%

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, tendo a presente versão sido revista e aprovada pelas investigadoras em questão.

6.1. Cristina Mota: Relatório

a) Actividades no âmbito do corte-e-costura:

- Colaboração na redacção e na preparação do poster relativo ao artigo Santos & Mota (2010)

- Apresentação do poster anterior na conferência LREC 2010 em Malta
- Implementação e teste de programas para contabilização do número e tipo de regras semânticas existentes, bem como do número de vezes em que são aplicadas aos corpos
- Esclarecimento de dúvidas de utilização do AC/DC e corte-e-costura
- Estudar formas de tornar o corte-e-costura mais eficiente
- Testes comparativos de eficiência entre o corte-e-costura e o vislcg3 de modo a avaliar a possibilidade de o motor de anotação passar a ser o vislcg3
- Melhorias e correcções ao programa corte-e-costura
- Criação do pacote de distribuição do corte-e-costura, incluindo
    - testes ao pacote fora das máquinas da Linguateca
    - redacção da documentação (Leiame, páginas html para o sítio dedicado ao corte-e-costura)
    - criação e teste de exemplos e materiais de teste (mantas de retalhos)

b) Anotação semântica: Escrita de regras de anotação da roupa

- revisão da palavra "malha" no CETEMPúblico
- identificação de novas palavras de roupa
- identificação de casos em que a roupa é ambígua em termos de categoria gramatical e em que a categoria pode estar mal atribuída (calças, calçado, vestido, uniforme, colar, brinco, saia)
- organização das regras por genéricas e específicas por corpo
- revisão da documentação sobre a anotação da roupa e esclarecimento de dúvidas com o Augusto Soares da Silva
- revisão de palavras anotadas como roupa no ENPCPUB
- início da revisão do CONDIV (revisão de casos em que a categoria gramatical de "meia", "calças" e "uniforme" foi mal atribuída)
- Detecção de problemas na anotação da cor
- Resolução de problemas de anotação (sejam correcções aos programas ou às regras)

c) Actividades no âmbito do Segundo HAREM:

- Colaboração na redacção e na preparação da apresentação do artigo Freitas e tal. (2010)
- Apresentação do artigo anterior na conferência LREC 2010 em Malta
- Criação e divulgação do novo pacote de distribuição da LÂMPADA:
    - revisão da documentação do pacote
    - alteração dos scripts de avaliação para acomodar a nova CD do ReRelEM
    - testes de execução da avaliação que incluem a nova CD do ReRelEM
- Revisão do artigo comparativo da avaliação do tempo no Primeiro e Segundo HAREM com vista à sua conclusão; interacção com a co-autora Paula Carvalho

d) Actividades no âmbito do ResPubliQA 2010:

- tradução de questões e delimitação e avaliação das respostas correctas
- avaliação das respostas dos sistemas participantes no ResPubliQA
- participação no artigo Peñas et al. (2010)

e) Outras actividades:

- Prospecção sobre a possibilidade de colaboração com o Nuno Cardoso na criação de uma DBpedia portuguesa
- Prospecção sobre a possiblidade de colaboração com o projecto Arquivo da Web para datar

páginas web

- Colaboração na preparação da apresentação e na apresentação do poster sobre o GikiCLEF (Santos et al. 2010) no LREC 2010

6.2. Rosário Silva: Relatório

a) Anotação semântica do AC/DC
Continuação da anotação da cor nos corpos do projecto AC/DC:
- Actualização das listas de cor, de grupos de cor e de regras a aplicar à anotação, assim como dos documentos sobre a anotação (Silva & Santos, 2010);
- Verificação da consistência das regras de ajuda à anotação nos corpos anotados com cor

– Revisão da anotação de vários corpos (CONDIV, CHAVE, CDHAREM, NATMINHO, NILC/SÃOCARLOS, MUSEU DA PESSOA, ANCIB, AMOSTRA, CONE, ECI-EBR, ECI-EE) e consequente correcção da mesma através da redacção de regras exclusivas de ajuda à anotação automática da cor;

b) Participação na avaliação conjunta ResPubliQA, desta feita traduzindo as perguntas indicadas;

c) Participação na análise de dados para o resumo a apresentar no congresso de línguas pluricênticas;

d) Passagem das informações necessárias sobre a anotação semântica da roupa nos corpos do projecto AC/DC à Cristina Mota;

e) Participação na escrita do artigo Santos et al. (2011).

6.3. Cláudia Freitas: Relatório

a) Actividades no âmbito do PAPEL

Em um primeiro momento, foram revisadas mais de 200 relações do PAPEL. 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, mencionam-se

-- participação no desenvolvimento da interface do VARRA
-- identificação e teste de regras para a extracção de relações semânticas
-- escrita de uma nova versão do manual voltado aos utilizadores/validadores do VARRA.
-- criação de dossiês de validação de relações semânticas, distribuídos aos validadores.
-- colaboração na redacção do resumo e artigo completo para o ELC, Freitas et al. (2011)
-- colaboração na preparação e apresentação do poster referente ao trabalho acima.
-- co-orientação, com a professora Violeta Quental, da aluna bolseira de iniciação científica da PUC-Rio Andrea Barreto.

b) Actividades no âmbito da Amazônia:

-- revisão da segmentação dos 1070 textos que compõem a Amazônia.
-- revisão dos URLs dos 1070 textos que compõem a Amazônia.
-- escrita do resumo e versão final do artigo Freitas (2011), "Amazônia: um corpus de blogs?".
-- preparação e apresentação do poster referente ao trabalho acima mencionado

c) Actividades no âmbito do Segundo HAREM - ReRelEM:

-- ajustes e correção de erros na CD do Segundo HAREM, disponibilizada no LAMPADA 2.0, no que se refere às relaçóes semânticas.

-- elaboracao de um glossário do ReRelEM, com exemplos das relações semânticas.

-- colaboração na escrita do artigo sobre o Segundo HAREM para o LREC, Freitas et al. (2010)

d) Outras actividades

-- colaboração na escrita do artigo Santos et al. (2011) sobre as diferenças entre as variantes

-- apresentação sobre a Linguateca para o grupo de PLN da PUC-Rio.

7. Publicações da Linguateca no período referente a este relatório

1) [Gonçalo Oliveira et al. 2010] Hugo Gonçalo Oliveira, Diana Santos & Paulo Gomes. "Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação". *Linguamática* **2.1** (2010), pp. 77-93. Nova versão, revista e aumentada, da publicação Gonçalo Oliveira et al (2009), no STIL 2009.
2) [Santos et al. 2010a] Diana Santos, Nuno Cardoso & Luís Miguel Cabral. "How geographic was GikiCLEF? A GIR-critical review". Apresentação na FCUL, Lisboa, 26 de Janeiro de 2010.
3) [Santos et al. 2010b] Diana Santos, Nuno Cardoso & Luís Miguel Cabral. "How geographical was GikiCLEF? A GIR-critical review". In *6th Workshop on Geographic Information Retrieval (GIR'10)* (Zurique, 18-19 Fevereiro de 2010).
4) [Ribeiro & Santos 2010] Fernando Ribeiro & Diana Santos. "Colaboração entre a Linguateca e o RCAAP: primeiros passos". Apresentação no *Encontro RCAAP* (Leiria, 22 de Março de 2010).
5) [Freitas 2010a] Cláudia Freitas. "Segundo HAREM, ReRelEM e LÂMPADA 2.0". Apresentação na PUC-Rio, Rio de Janeiro, Abril de 2010).
6) [Freitas et al. 2010a] 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 Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)* (Valletta, Malta, 17-23 de Maio de 2010), European Language Resources Association, pp. 3630-3637.
7) [Santos et al. 2010b] Diana Santos, Luís Miguel Cabral, Corina Forascu, Pamela Forner, Fredric Gey, Katrin Lamm, Thomas Mandl, Petya Osenova, Anselmo Peñas, Alvaro Rodrigo, Julia Schulz, Yvonne Skalban & Erik Tjong Kim Sang. "GikiCLEF: Crosscultural issues in multilingual information access". In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)* (Valletta, Malta, 17-23 de Maio de 2010), European Language Resources Association, pp. 2346-2353.
8) [Santos et al. 2010c] Diana Santos, Luís Miguel Cabral, Pamela Forner, Corina Forascu, Fredric Gey, Katrin Lamm, Thomas Mandl, Petya Osenova, Anselmo Peñas, Alvaro Rodrigo, Julia Schulz, Yvonne Skalban, Erik Tjong Kim Sang & Nuno Cardoso. "GikiCLEF: Crosscultural issues in multilingual information access". Poster na *International Conference on Language Resources and Evaluation (LREC 2010)* (Valletta, Malta, 17-23 de Maio de 2010).
9) [Santos & Mota 2010] Diana Santos & Cristina Mota. "Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora". In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)* (Valletta, Malta, 17-23 de Maio de 2010), European Language Resources Association, pp. 1437-1444.

10) [Santos & Ribeiro 2010] Diana Santos & Fernando Ribeiro. "Estudando os autores: Trabalho referente à colaboração com o RCAAP". Linguateca, FCCN. 6 de Agosto de 2010.
11) [Santos & Finatto 2010] Diana Santos & Maria José Bocorny Finatto. "Words and their secrets". *ESSLLI 2010* (Dinamarca, Agosto de 2010). Curso de uma semana em Escola de Verão internacional.
12) [Cardoso 2010] Nuno Cardoso. "GikiCLEF topics and Wikipedia articles: Did they blend?". In Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas & Giovanna Roda (eds.), *Multilingual Information Access Evaluation, VOL I* Setembro de 2010, Springer.
13) [Santos & Cabral 2010] Diana Santos & Luís Miguel Cabral. "GikiCLEF : Expectations and lessons learned". In Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas & Giovanna Roda (eds.), *Multilingual Information Access Evaluation, VOL I* Setembro de 2010, Springer, pp. 212-222.
14) [Peñas et al. 2010] Anselmo Peñas, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, Corina Forascu & Cristina Mota. "Overview of ResPubliQA 2010: Question Answering Evaluation over European Legislation". In *ResPubliQA - Multilingual Question Answering at CLEF 2010 (QA@CLEF 2010 - ResPubliQA)* (Padua, Itália, 20-23 Setembro).
15) [Santos et al. 2010d] Diana Santos, Rosário Silva, Cláudia Freitas & Augusto Soares da Silva. Pluralidades na cor: contrastando a língua do Brasil e de Portugal". Apresentação no congresso *Línguas Pluricêntricas: Variação Linguística e Dimensões Sociocognitivas* (Braga, Portugal, 15-17 de Setembro de 2010).
16) [Santos 2010] Diana Santos. "A rede de tradução e como operacionalizá-la a partir de corpos paralelos: primeiros passos". Apresentação no *I Workshop Per-Fide: Construção, exploração e aplicação de Corpora Paralelos (WPerfide1)* (Braga, Portugal, 16-17 de Setembro de 2010).
17) [Freitas et al. 2010b] Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira & Violeta Quental. "VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC". Poster no *ELC2010* (Porto Alegre, 8-9 de Outubro de 2010).
18) [Freitas 2010b] Cláudia Freitas. "Amazônia: um corpus de blogs?". Poster no *ELC2010* (Porto Alegre, 8-9 de Outubro de 2010).
19) [Santos et al. 2010e] Diana Santos, Anabela Barreiro, Cláudia Freitas, Hugo Gonçalo Oliveira, José Carlos Medeiros, Luís Costa, Paulo Gomes & Rosário Silva. "Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL". In A. M. Brito, F. Silva, J. Veloso & A. Fiéis (eds.), *Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística*. APL, 2010, pp. 681-700.
20) [Silva & Santos 2010] 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 constante revisão, última versão: 26 de Dezembro de 2010.
21) [Santos, Silva & Mota 2010] Diana Santos, Augusto Soares da Silva & Cristina Mota. "Guarda-fatos: notas sobre a anotação do campo semântico do vestuário em português". Primeira edição: 26 de Outubro de 2009. Em constante revisão, última versão: 13 de Julho de 2010.

No prelo

22) [Teixeira et al. no prelo] 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.), *Caminhos na Linguística de Corpus*. Mercado de Letras.
23) [Santos no prelo] Diana Santos. "Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties". *OSLa: Oslo Studies in Language* (2011). ISSN: 18909639.

Enviados para apreciação

24) [Maia & Santos 2011] Belinda Maia & Diana Santos. "Who is afraid of what?: fear in English and Portuguese". Enviado a *ICAME2011*.
25) [Santos et al. 2011] Diana Santos, Stella E. O. Tagnin & Elisa Duarte Teixeira. "Colours, clothing and food in CorTrad: why corpus-based translation studies are revealing". Enviado a *ICAME2011*.

Em preparação

26) [Santos et al. 2011] Diana Santos, Rosário Silva, Cláudia Freitas & Augusto Soares da Silva. Pluralidades na cor: contrastando a língua do Brasil e de Portugal".
27) [Freitas et al. 2011] Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira & Violeta Quental. "VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC".
28) [Freitas 2011] Cláudia Freitas. "Amazônia: um corpus de blogs?".

Índice