Relatório da Linguateca relativo ao ano de 2011

Diana Santos
Dezembro de 2011

A Linguateca foi reestruturada em 2011 para funcionar como um projeto não só de manutenção e disponibilização de recursos e gerador de avaliações conjuntas na área do processamento computacional da língua portuguesa, mas também para iniciar iniciativas de maior impacto a nível da sociedade, quer através da implementação de iniciativas estruturantes como através da colaboração com outros projetos mais aplicados. A sua continuação foi planeada ao abrigo do financiamento como projeto especial da Fundação para a Ciência e a Tecnologia, pelo lapso de três anos (2011-2013), condicionada à aprovação, cada ano, com base no conseguido no(s) ano(s) anterior(es). Este é o relatório do primeiro ano.

Em termos administrativos, e no que se refere à contratação de recursos humanos, aquilo que nos foi concedido ficou infelizmente bastante aquém do planeado e aprovado. Em vez de 4,95 pessoas no primeiro ano, apenas pudemos contar com os seguintes colaboradores, totalizando 3,15 pessoas num ano, correspondendo portanto a 36% menos do que o inicialmente planeado.

A equipa constou pois dos seguintes colaboradores:

Contratados pela FCCN: Fernando Ribeiro a 100% e Diana Santos a 20% o ano todo.
Em regime de contrato de prestação de serviços

- Rosário Silva a 25% o ano todo,
- Cristina Mota a 100% desde fevereiro de 2011,
- Cláudia Freitas a 30% de janeiro a junho de 2011 e 50% a partir de julho de 2011,
- Luís Costa a 100% desde 8 de agosto de 2011,
- Hernâni Costa a 50% desde setembro de 2011.

Além disso, tivemos a colaboração, graciosa ou através de fontes de financiamento externas ou de colaborações com outras instituições, dos seguintes investigadores ligados à rede da Linguateca: Alberto Simões, Nuno Cardoso, Hugo Gonçalo Oliveira, Alice Gonçalves, Marcin Wlodek e Paulo Rocha, como é parcialmente refletido nas publicações. Finalmente, o trabalho com o CorTrad beneficiou naturalmente da numerosa equipa do COMET, liderada por Stella Tagnin na USP.

Trabalho realizado

Esta nova fase da Linguateca, conforme descrito na proposta enviada à FCT, desenvolve-se em três sentidos distintos, que não são naturalmente estanques e que faz sentido desenvolver em conjunto, conforme defendido na proposta aprovada.

A. Desenvolvimento de uma gramática descritiva baseada em métodos empíricos para o português usando os recursos linguísticos da Linguateca e os recursos computacionais da Universidade de Oslo.

B. Organização de uma infraestrutura para estudos (linguístico-)culturais da lusofonia e seu contraste com outras línguas e culturas

C. Estudo e melhoria do RCAAP (Repositório Científico de Acesso Aberto de Portugal) **no que se refere às particularidades da língua portuguesa**

Além disso, a Linguateca continua a manter a infra-estrutura de serviço aos múltiplos utilizadores, com adições ao catálogo, ao fórum, ao catálogo de publicações, e estatísticas mensais de acesso aos nossos serviços e recursos, apoiando ativamente os seus utilizadores, ao que corresponde também

um esforço de documentação relativamente apreciável, e que descreveremos como:

D. Continuação de apoio e desenvolvimento dos recursos da Linguateca.

Passamos a detalhar o trabalho realizado em cada uma das vertentes:

Trabalho desenvolvido na vertente A

A construção de materiais didáticos de gramática baseados em corpos foi iniciada através do desenvolvimento do Ensinador[1] (Simões & Santos, 2011)

Foram desenvolvidas outras ferramentas ou funcionalidades para melhorar a interação com os corpos em português, ainda em fase de protótipo, como o Comparador e o Distribuidor.

O foco em descrições contrastivas (com o inglês e outras línguas), o que levou à continuação do desenvolvimento de corpos bilingues e de tradução, em particular o CorTrad[2] e o PoNTE[3].

1. Anotação e revisão dos campos semânticos da cor e roupa no CorTrad – revisão apenas no português (Santos et al., 2011, 2012a,b).
2. Revisão, por bolseiros da Universidade de São Paulo, do alinhamento dos subcorpos CorTrad jornalístico e literário, com a consequente instalação de novas versões no sítio do projeto.
3. Início da criação do CorTrad resumos de teses (parada neste momento do lado brasileiro) e contatos para uma versão para ensino de português como língua estrangeira com muitas traduções
4. Implementação da primeira fase do projeto PoNTE (com co-financiamento da Universidade de Oslo), de traduções entre o português e o norueguês.

Embora o objetivo último, nomeadamente a construção de uma gramática baseada em corpos, não tenha ainda sido iniciado, todos estes passos são facilitadores desse fim, assim como a melhoria significativa dos próprios corpos, relatada na vertente seguinte.

Trabalho desenvolvido na vertente B

Esta linha dividiu-se naturalmente em duas áreas de atividade: o desenvolvimento da própria infra-estrutura e conteúdo, à volta do AC/DC[4], e a organização do Págico, que foi ao que comparativamente foi dedicado maior esforço pela equipa, devido ao compromisso com os participantes e os prazos apertados.

1) Melhoria significativa de vários corpos no AC/DC e novos serviços à volta da semântica da língua portuguesa
   a) O conteúdo do corpo VERCIAL foi significativamente melhorado com a remoção de texto em língua estrangeira e uma melhor separação de frases em textos de teatro e poesia.
   b) Ao corpo CONDIV foi efetuada a remoção de textos demasiado pequenos e vários corpos receberam melhor segmentação.
   c) Foi efetuada a revisão da anotação da cor num número significativo de corpos, e melhoria da documentação associada (Silva e Santos, em constante atualização, Freitas et al., 2011, Freitas, 2011).

[1] http://www.linguateca.pt/Ensinador
[2] http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html
[3] http://dinis.linguateca.pt/dispara/ponte/
[4] http://www.linguateca.pt/ACDC/

d) Foi executado um estudo da homografia no campo semântico da roupa e procedeu-se à revisão da sua anotação em alguns corpos (Santos, Soares da Silva & Mota, em constante atualização).
e) Criação de anotação em relação ao campo semântico do medo (Maia & Santos, 2011) e sua aplicação (automática) a todos os corpos do AC/DC.
f) Início da anotação com o REMBRANDT[5] das páginas (em português) comuns entre as wikipédias portuguesa e norueguesas, assim como anotação completa da coleção CHAVE[6] (Cardoso & Santos, 2012).
g) Interligação do PAPEL e do TeP[7] com o AC/DC, através da adição de campos contendo os sinónimos, os antónimos, e os hiperónimos de cada palavra a todos os corpos.
h) Desenvolvimento de um novo Folheador[8] (Gonçalo Oliveira et al., 2012) para um conjunto de recursos semânticos públicos para o português, e que permite também a invocação do VARRA[9] e do AC/DC.
i) Melhoria das funcionalidades do programa corte-e-costura[10], de apoio a anotação semântica dos corpos.

2) Organização da avaliação conjunta Págico[11], em progresso
a) Sua divulgação atempada, em português e inglês, com a construção de um sítio dedicado a esta avaliação e um folheto de divulgação;
b) Re-instalação do sistema SIGA para gestão dessa avaliação,
c) Considerável melhoria desse sistema, no que se refere à possibilidade de responder aos tópicos interativamente e procurar justificações;
d) Preparação da coleção do Págico processando a wikipédia;
e) Criação de 150 tópicos diversificados sobre assuntos de cultura lusófona presentes na wikipédia em português;
f) Avaliação das 52882 respostas (50184 automáticas e 2698 humanas), em progresso;
g) Extensa comunicação com os participantes e esclarecimento de dúvidas;
h) Escrita de artigos ou resumos alargados de divulgação (Costa et al., 2012, Mota et al., 2012);
i) Início da organização de uma edição especial da revista *Linguamática* dedicada ao Págico, que deverá sair por altura do encontro final do Págico em 21 de abril de 2012 no PROPOR em Coimbra.

Trabalho desenvolvido na vertente C

1. Consolidado o processamento mensal, com a correspondente operacionalização da noção de sessão e a criação de estatísticas robustas, num sítio dedicado[12], veja-se Santos & Ribeiro (2011);
2. Primeiros passos na construção de uma infra-estrutura genérica de estudo de utilizadores e sessões, aplicada também ao AC/DC;
3. Estudos detalhados sobre o acesso ao RCAAP através do meta-repositório (Santos & Ribeiro, 2012a, Ribeiro & Santos, 2012b);
4. Recolha de dois conjuntos de publicações com intuitos específicos, para estudar mecanismos de citação de referências: as publicações com texto público constantes do catálogo de publicações da Linguateca, e as citações a um autor específico, provenientes do Google

[5] http://xldb.di.fc.ul.pt/Rembrandt/
[6] http://www.linguateca.pt/CHAVE/
[7] http://www.nilc.icmc.usp.br/tep2/
[8] http://www.linguateca.pt/Folheador/
[9] http://www.linguateca.pt/VARRA/
[10] http://www.linguateca.pt/acesso/corte-e-costura/
[11] http://www.linguateca.pt/Pagico/
[12] http://www.linguateca.pt/colabRCAAP/

Scholar;
5. Limpeza e melhoria do catálogo de publicações da Linguateca, com consequente purga de entradas repetidas e inserção de novas categorias no SUPeRB[13].

Trabalho desenvolvido na vertente D

Os seguintes trabalhos adicionais – embora na suma maioria não planeados – são também dignos de nota, por indicarem que a atividade da Linguateca como centro de recursos e de informação sobre o processamento do português é uma realidade viva e dinâmica:

1. a reanotação da coleção HAREM[14] em relação às entidades temporais (Mota & Carvalho, 2011);
2. a migração da GeoNet-PT[15] para as máquinas da Linguateca;
3. a documentação aturada das relações do PAPEL no seu sítio;
4. o recálculo melhorado das listas de frequências do AC/DC;
5. o dimensionamento de uma nova máquina, virtual, para alojar os serviços da Linguateca e a sua população e entrada em funcionamento;
6. a criação do PAPEL 3.0[16] e sua disponibilização;
7. a renovação total do sítio do projeto VARRA[17], com a separação entre duas tarefas diferentes: validar triplos, por um lado, e descobrir padrões, por outro;
8. a racionalização do sítio e da distribuição da WPT[18];
9. o apoio avançado a utilizadores dos serviços AC/DC, Floresta[19] e SAHARA[20] (sistema de avaliação automático do HAREM);
10. a resposta a pedidos de esclarecimento ou aconselhamento sobre a) "POS-taggers", b) reconhecimento de entidades mencionadas, c) recolha de informação geográfica, e d) análise sintática dependencial, para o português.

Em resumo, continuámos a atividade normal da Linguateca como fornecedora de recursos e incentivadora do seu uso, com aliás a seguinte tabela, com o acesso a diferentes recursos em 2011, e cerca de 2 milhões e meio de acessos ao nosso sítio, eloquentemente testemunha:

Recurso	**Levantamentos ou acessos**
CETEMPúblico[21]	58
CETENFolha[22]	60
CHAVE	20
PAPEL	66
Esfinge[23]	32
Floresta	110
GIRA[24]	13

[13] http://www.linguateca.pt/SUPeRB/
[14] http://www.linguateca.pt/HAREM/
[15] http://www.linguateca.pt/geonetpt/
[16] http://www.linguateca.pt/PAPEL/
[17] http://www.linguateca.pt/VARRA/
[18] http://www.linguateca.pt/WPT/
[19] http://www.linguateca.pt/Floresta/
[20] http://www.linguateca.pt/SAHARA/
[21] http://www.linguateca.pt/CETEMPublico/
[22] http://www.linguateca.pt/CETENFolha/
[23] http://www.linguateca.pt/Esfinge/

LÂMPADA[25]	67
Outros	29

Para uma visão mais abrangente dos acessos e interação dos nossos utilizadores com o sítio da Linguateca, veja-se de qualquer maneira a nossa página de estatísticas, atualizada mensalmente.

Comentário geral

Este primeiro ano lançou as bases para uma infra-estrutura capaz de fundamentar uma gramática e de produzir estudos contrastivos em larga escala, demonstrando alguns sistemas e serviços inovadores. O grosso do trabalho e da atividade centrou-se contudo na organização do Págico, que pela primeira vez reuniu pessoas e sistemas automáticos na procura de informação sobre a lusofonia, e que será apresentado, e criticado, publicamente durante o encontro satélite do PROPOR 2012 em Coimbra.

Não será contudo de desprezar o primeiro estudo crítico da interação de utilizadores com um serviço público em português na área da procura de publicações, pese embora a redução no financiamento e portanto no pessoal apto a realizar esse trabalho.

Pensamos, aliás, que ao nível do púbico em geral continuámos a dar o apoio e os recursos a que o habituámos, e que faz da Linguateca um ator importante, e sobretudo um serviço com que as pessoas contam, na área.

Segue-se a lista de publicações, e os relatórios individuais de todos os contratados, levemente editados e resumidos por mim.

[24] http://www.linguateca.pt/GikiCLEF/GIRA/
[25] http://www.linguateca.pt/HAREM/PacoteRecursosSegundoHAREM.zip

Publicações

Mais uma vez, tentámos documentar e divulgar o trabalho realizado na Linguateca através de uma atividade de publicação contínua. Neste relatório separamos o seu resultado de acordo com o estatuto das publicações (publicadas, no prelo, enviadas para apreciação, e em preparação).

Publicações no período a que se refere o presente relatório

1. Cristina Mota & Paula Carvalho. "O passar do TEMPO no HAREM". *Linguamática* **3.1** (2011), pp. 45-58.
2. Diana Santos & Fernando Ribeiro. "Estudando os nomes dos autores no RCAAP: relatório do primeiro ano". Relatório FCCN, 4 de junho de 2011.
3. Diana Santos, Rosário Silva & Cláudia Freitas. "Pluralidades na cor: contrastando a língua do Brasil e de Portugal". In Augusto Soares da Silva, Amadeu Torres & Miguel Gonçalves (eds.), *Línguas Pluricêntricas: Variação Linguística e Dimensões Sociocognitivas*. Braga : Aletheia, Publicações da Faculdade de Filosofia da Universidade Católica Portuguesa, 2011, pp. 535-552.
4. Diana Santos. "Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties". *OSLa: Oslo Studies in Language* **3.2** (2011), pp. 113-128. ISSN: 18909639. Volume edited by J.B.Johannessen, Language variation infrastructure.
5. Alberto Simões & Diana Santos. "Ensinador: corpus-based Portuguese grammar exercises". *Procesamiento del Lenguaje Natural* **47** (2011), pp. 301-309.
6. Hernâni Costa, "O desenho do novo Folheador". Relatório técnico, Linguateca, 2011
7. Rosário Silva & Diana Santos. "Arco-íris: notas sobre a anotação do campo semântico da cor em português". Em constante atualização. Primeira edição: 25 de junho de 2009.
8. Diana Santos, Augusto Soares da Silva & Cristina Mota. "Guarda-fatos: notas sobre a anotação do campo semântico do vestuário em português". Em constante atualização. Primeira edição: 26 de outubro de 2009.
9. Cláudia Freitas, Diana Santos & Alice Gonçalves . “Perguntas já respondidas sobre o AC/DC: desde como começar até uso complexo de funcionalidades poderosas”. Em constante atualização. Primeira edição: 15 de novembro de 2011.

Apresentações

10. Belinda Maia & Diana Santos. "Who is afraid of. what?: Fear in English and Portuguese". In *ICAME2011*, Oslo, 2 de junho de 2011.
11. Diana Santos, Stella E. O. Tagnin & Elisa Duarte Teixeira. "Colours, clothing and food in CorTrad: why corpus-based translation studies are revealing". In *ICAME2011*, Oslo, 2 de junho de 2011.
12. Diana Santos. "Compreensão de linguagem natural: voltando à carga". Universidade de Aveiro, 18 de julho de 2011.
13. Diana Santos. "Translation and categorization". Universidade de Oslo, 29 de setembro de 2011.
14. Diana Santos. "À procura do tempo perdido / In search of the lost time/tense". Universidade de Oslo, 6 de outubro de 2011.
15. Stella E. O. Tagnin. "CorTrad: um corpus para ajudar aprendizes de tradução a obterem um texto natural". In *ENCULT - II Encontro Nacional de Cultura e Tradução*, UFPB, João Pessoa, 5-7 de outubro de 2011.
16. Cláudia Freitas, Diana Santos & Rosário Silva. “Corpos e cores: colorindo a descrição da Língua Portuguesa”. *X Encontro de Linguística de Corpus*. Belo Horizonte, Brasil, 11-12 de novembro de 2011.

17. Cláudia Freitas. “Os corpos da Linguateca na prática” (mini-curso). *V Escola Brasileira de Linguística Computacional/ X Encontro de Linguística de Corpus*. Belo Horizonte, Brasil, 11-12 de novembro de 2011.

No prelo: Artigos enviados para publicação

18. Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira & Violeta Quental. "VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC". In *Anais do ELC2010*, 2012, no prelo.
19. Cláudia Freitas & Diana Santos. "Blogs, Amazônia e a Floresta Sintá(c)tica: um corpus de um novo gênero?". In A*nais do ELC2010*, 2012, no prelo.
20. 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, 2010, no prelo.

Em preparação: Artigos aceites para publicação em 2012

21. Diana Santos. "Corpora at Linguateca". In Tony Berber Sardinha & Telma São Bento Ferreira (eds.), *Working with Portuguese corpora*. Continuum, 2012.
22. Diana Santos. "The next step for the translation network”. In Diana Santos, Krister Lindén & Wanjiku Ng’ang’a (eds.), *Shall we play the Festschrift game? Essays on the occasion of Lauri Carlson's 60th birthday.* Springer, 2012.
23. Diana Santos. “Porquê o Págico?” in *Linguamática* 4.1, editada por Mota et al. (2012)
24. Cláudia Freitas. “A lusofonia na wikipédia em 150 tópicos”, *Linguamática* 4.1, editada por Mota et al. (2012).
25. Alberto Simões, Cristina Mota & Luís Costa. “A wikipédia em português no Págico: adaptação e avaliação”, in *Linguamática* 4.1, editada por Mota et al. (2012)
26. Luís Costa e Cristina Mota. “A transformação do SIGA para o Págico”, in *Linguamática* 4.1, editada por Mota et al. (2012)
27. Cristina Mota. “Resultados págicos: participação, resultados e recursos”, in *Linguamática* 4.1, editada por Mota et al. (2012)
28. Cristina Mota, Cláudia Freitas & Luís Costa. “O que é uma resposta? Notas de uns avaliadores estafados”, in *Linguamática* 4.1, editada por Mota et al. (2012)
29. Cristina Mota & Diana Santos. “Balanço e contributos para a definição do próximo Págico”, n *Linguamática* 4.1, editada por Mota et al. (2012) .

Em apreciação: Artigos enviados para apreciação

30. Luís Costa, Cristina Mota & Diana Santos. "SIGA, a Management System to Support the Organization of Information Retrieval Evaluations".
31. Hugo Gonçalo Oliveira, Hernâni Costa & Diana Santos. "Folheador: browsing through Portuguese semantic relations".
32. Nuno Cardoso & Diana Santos. "Where are we in CHAVE?".
33. Diana Santos & Fernando Ribeiro. "Uma incursão pelo universo das publicações em Portugal".

Em apreciação: Resumos enviados para apreciação

34. Fernando Ribeiro & Diana Santos. "Studying the names of authors in RCAAP, a national repository of open source publications".
35. Cristina Mota, Alberto Simões, Cláudia Freitas, Luís Costa & Diana Santos. “Págico: EvaluatingWikipedia-based information retrieval in Portuguese”.
36. Diana Santos, Stella E. O. Tagnin, Elisa Duarte Teixeira. “CorTrad and Portuguese-English translation studies: colours and clothing”.

37. Belinda Maia & Diana Santos. “Who’s afraid of …… what?” – in English and Portuguese.
38. Diana Santos, Stella E. O. Tagnin, Elisa Duarte Teixeira. “CorTrad search features and translation studies: a pilot study on colours, clothing and food domains”.

Relatório de Cristina Mota

Atividades no âmbito do Págico

- Instalação e adaptação (do GikiCLEF para o Págico) do sistema SIGA, nos seguintes pontos
    - avaliação baseada também nas justificações adicionadas pelos donos dos tópicos
    - avaliação de respostas de participantes humanos
    - diversas alterações à interface por o Págico só lidar com uma língua (português)
  - Tradução do SIGA de inglês para português
    - Revisão e melhoria da documentação do SIGA
    - Nova interface de gestão e de avaliação devido à participação humana
    - Adição de temas e supertemas aos tópicos
- Gestão e manutenção genérica do SIGA
- Geração/conversão de páginas da Wikipédia portuguesa para XML, para incorporação no SIGA, e sua disponibilização aos participantes (várias versões)
- Instalação da nova coleção de documentos no SIGA
- Instalação de tópicos de exemplo, e participação na criação dos tópicos finais
- Gestão e manutenção genérica do sítio do Págico
- Participação na avaliação das respostas do Págico
- Criação, teste e divulgação de recursos do Págico
- Colaboração na organização do Págico
    - Participação na escrita de dois artigos em inglês relacionados com o Págico (um resumo alargado enviado para apreciação ao LREC 2012, e um artigo completo ao PROPOR 2012)
    - Participação na escrita de quatro artigos sobre o Págico em português para a edição especial da Linguamática sobre o mesmo (Simões & Mota, Costa et al, Mota, Mota & Santos)
    - Como editora principal da edição especial, trabalho de chamada de artigos e definição do volume.

Atividades no âmbito da anotação de corpos

- Recauchutamento dos serviços na rede dos projetos Varra e AC/DC para usar o novo Corpus Workbench e respectivos módulos de perl
- Implementação do Ensinador, uma ferramenta para a geração de exercícios para aprendizagem da língua com base em corpos
- Implementação do Distribuidor, uma ferramenta para o cálculo de distribuições de diferentes atributos em corpos, a nível de protótipo
- Implementação do Comparador, uma ferramenta para a comparação de pesquisas entre corpos, a nível de protótipo
- Implementação de uma ferramenta para a anotação de informação semântico-lexical em corpos, com base no PAPEL e no TeP (anotação de sinónimos, antónimos e hiperónimos)
- Escrita de um relatório técnico interno sobre as cinco ferramentas acima mencionadas
- Implementação de programas para contagem de homografias com palavras de roupa
- Actualização do Guarda-fatos (Santos et al., em constante atualização), com

    - informações sobre a anotação da roupa feita no âmbito do contrato anterior
    - informações sobre a hierarquia de classificação da roupa
    - contagens das homografias com palavras de roupa existentes nos corpos do AC/DC
- Anotação da prosa nos corpos do Vercial
- Atualização do corte-e-costura para poder aceitar outros elementos estruturais além de mwe
- Correção de alguns problemas detetados no corte-e-costura

Atividades no âmbito do HAREM

- Conclusão e publicação na revista Linguamática do artigo comparativo da avaliação do tempo no Primeiro e Segundo HAREM, intitulado "O passar do TEMPO no HAREM" (Mota e Carvalho, 2011)
- Revisão da reanotação das entidades classificadas como TEMPO na colecção dourada do Primeiro HAREM de acordo com as directivas do Segundo HAREM
- Reanotação das entidades classificadas como TEMPO na colecção dourada do Segundo HAREM de acordo com as directivas do Primeiro HAREM
- criação da página que apresenta sucintamente e disponibiliza as CD do Primeiro e Segundo HAREM reanotadas, e atualização da página do HAREM
- disponibilização e divulgação em listas nacionais e internacionais das CD do Primeiro e Segundo HAREM

Atividades de desenvolvimento e melhoria de outros recursos

- apoio a um utilizador da Lâmpada e do SAHARA
- esclarecimento de dúvidas sobre analisadores morfossintáticos de português

Relatório de Cláudia de Freitas

Atividades no âmbito do projeto VARRA

- Preparação e disponibilização, na página do VARRA, de mais dossiês de validação de relações semânticas entre palavras.
- Colaboração no redesenho da página do VARRA, que foi divida em três páginas: a página inicial propriamente, o VARRA para validar relações e o VARRA para descobrir relações.
- Formalização de regras referentes a padrões lexicais para extração de relações semânticas para inclusão na página do VARRA.
- Envio da versão final do artigo "VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC". (Freitas et al. 2011).
- Como decorrência da melhoria da página e documentação do VARRA, foi também disponibilizada uma página com a descrição das relações semânticas presentes no PAPEL, projeto ao qual o VARRA está diretamente relacionado.
- Discussão com a equipe do PAPEL sobre determinadas regras e relações semânticas obtidas automaticamente.

Atividades no âmbito da Floresta Sintá(c)tica / Amazônia

- Estudo das características linguísticas da Amazônia, tendo em vista o estudo de marcadores específicos de blogues, contrastando-os especificamente com marcadores presentes no corpo do Museu da Pessoa e em seções específicas do AC/DC.
- Envio da versão final do artigo "Blogs, Amazônia e a Floresta Sintá(c)tica: um corpus de um novo gênero?" (Freitas & Santos, 2011)

Atividades no âmbito do projeto AC/DC

- Seleção de perguntas relacionadas ao AC/DC de um arquivo com toda a correspondência eletrônica enviada para a Linguateca, para substituição da página de exemplos por uma página de "perguntas já respondidas" sobre buscas linguísticas no sistema. O referido documento (Freitas, Santos e Gonçalves, 2011) está disponível em formato html e pdf.
- Correção da segmentação dos textos de prosa do corpo Vercial.
- Envio da versão final do artigo "Pluralidades na cor: contrastando a língua do Brasil e de Portugal" (Santos et al. 2011)
- Escrita do resumo Freitas & Santos (2011), "Corpos e cores: colorindo a descrição da Língua Portuguesa" enviado ao ELC 2011.
- Preparação e elaboração de um mini-curso sobre o AC/DC, ministrado durante a V Escola Brasileira de Linguística Computacional (Freitas, 2011).
- Preparação de relatório com pontos a serem discutidos relativos à anotação das cores.
- Continuação da exploração do campo semântico das cores no AC/DC, tendo em vista a apresentação feita no X Encontro de Linguística de Corpus do trabalho "Corpos e cores: colorindo a descrição da Língua Portuguesa" (Freitas, Santos e Silva, 2011).

Atividades no âmbito do Págico

- Leitura do material relativo ao GikiCLEF para familiarização com o tipo de avaliação realizada no Págico;
- Criação e revisão de tópicos / perguntas para o Págico.
- Colaboração na avaliação das respostas enviadas ao Págico
- Colaboração na escrita do resumo alargado "Págico: EvaluatingWikipedia-based

information retrieval in Portuguese" (Mota et al 2011), enviado para apreciação ao LREC

- Escrita dos artigos "A lusofonia na wikipédia em 150 tópicos" (Freitas 2012) e "O que é uma resposta? Notas de uns avaliadores estafados" (Mota et al. 2012) para a edição especial da Linguamática dedicada ao Págico

Relatório de Rosário Silva

- Verificação e melhoria dos ficheiros referentes aos Grupos e às Classes da cor.
- Correcção de alguns erros detectados na classe cor:original do CONDIV
- Revisão do corpo NILC/São Carlos e redação de regras exclusivas para melhorar e corrigir a anotação deste corpus. Optou-se pela comparação de duas versões anotadas do corpus NILC/São Carlos numa tentativa de tornar as regras e respetiva aplicação mais eficientes.
- Revisão do corpus Vercial e redação de regras exclusivas para melhorar e corrigir a anotação deste corpus.
- Colocação de etiquetas de anotação em um terço dos textos de poesia do corpus Vercial.
- Revisão do corpus NatMinho e redação de regras exclusivas para melhorar e corrigir a anotação deste corpus (não concluído, ainda falta uma nova passagem).
- Revisão do corpus CETEMPúblico e redação de regras exclusivas para melhorar e corrigir a anotação deste corpo (em progresso).
- Revisão de vários outros corpos mais pequenos.
- Apoio em relação à anotação da cor do projeto Cortrad.
- Tratamento dos diminutivos e superlativos no que se refere às cores.
- Atualização do documento Rosário Silva & Diana Santos. "Arco-íris: notas sobre a anotação do campo semântico da cor em português".
- Participação no artigo Diana Santos, Rosário Silva & Cláudia Freitas. "Pluralidades na cor: contrastando a língua do Brasil e de Portugal". In Augusto Soares da Silva, Amadeu Torres & Miguel Gonçalves (eds.), *Línguas Pluricêntricas: Variação Linguística e Dimensões Sociocognitivas*. Braga: Aletheia, Publicações da Faculdade de Filosofia da Universidade Católica Portuguesa, 2011, pp. 555-572.
- Participação no artigo Cláudia Freitas, Diana Santos & Rosário Silva. "Corpos e cores: colorindo a descrição da língua portuguesa". In *ELC2011* (Minas Gerais, Brasil, 11-12 de Novembro de 2011).

Relatório de Luís Costa

No contexto do contrato de prestação de serviços no quadro do projecto Linguateca vigente entre 8 de Agosto e 31 de Dezembro de 2011 dediquei a totalidade do tempo à co-organização do Págico visto que isso foi considerado uma prioridade da Linguateca, desempenhando as seguintes tarefas:

- Participação na organização da avaliação conjunta Págico, o que me levou a seguir e participar nas discussões da organização relativas avários aspectos da avaliação.
- Instalação local do sistema SIGA, estudo da documentação e familiarização com o código do mesmo.
- Extensão da estrutura da base de dados usada pelo SIGA para suportaros novos requisitos do Págico.
- Extensão da interface de criação de tópicos do SIGA de forma a permitir armazenar justificações para as respostas.
- Criação de várias versões de uma interface para suportar a participação humana no Págico.
- Participação na escrita de dois artigos em inglês relacionados com o Págico (um enviado para apreciação ao LREC 2012, outro ao PROPOR 2012).
- Cálculo de estatísticas sobre a participação humana no Págico.
- Avaliação de respostas dos participantes no Págico.
- Avaliação da wikipédia com base nos tópicos de Págico
- Participação na escrita de dois artigos sobre o Págico em português para a edição especial da Linguamática sobre o mesmo

Relatório de Hernâni Costa

- Desenvolvimento e teste de um novo sistema de manuseamento e apresentação de relações semânticas em português, Folheador, com o conteúdo de vários recursos públicos, com ligação a outros serviços da Linguateca, nomeadamente o AC/DC e o VARRA, e com o cálculo de valores de confiança.
- Desenho da interface com atenção a problemas de usabilidade.
- Estudo de vários sistemas de visualização gráfica e escolha de um para acoplar ao Folheador.
- Participação na escrita de um artigo enviado para apreciação: Hugo Gonçalo Oliveira, Hernâni Costa & Diana Santos. "Folheador: browsing through Portuguese semantic relations".
- Escrita de documentação técnica do sistema desenvolvido: Hernâni Costa, "O desenho do novo Folheador". Relatório técnico, Linguateca, 2011.