Relatório da Linguateca
de 15 de Maio de 2004 a 14 de Maio de 2005

Diana Santos
2 de Junho de 2005

Referência do Projecto: POSI/PLP/43931/2001
Título do Projecto: Centro de Recursos Distribuído para o Processamento Computacional da Língua Portuguesa

Neste relatório faz-se uma breve sinopse das actividades desenvolvidas no âmbito da Linguateca, em todos os seus pólos. Sendo um projecto distribuído, muitas das actividades são feitas em comum. Contudo, por nos encontrarmos sensivelmente a meio do projecto, apresentamos também um relatório relativo a cada pólo, além dos relatórios individuais dos bolseiros.

Com no relatório anterior, a maior parte da informação e resultados encontram-se disponíveis na rede (WWW), donde não se repetem aqui, apresentando apenas a indicação da localização desse conteúdo na rede.

Estrutura do relatório

1. Panorâmica global

1.1 Trabalho efectuado

Organização do livro sobre avaliação conjunta (24 capítulos, 430 páginas)

Redacção de 14 artigos total ou parcialmente no âmbito da Linguateca
Revisão cruzada dos vários artigos
Formatação e tratamento aturado da bibliografia
Contacto com editores e envio do livro final à Caminho
http://acdc.linguateca.pt/aval_conjunta/LivroAvalon/

Continuação da organização da participação do português no CLEF http://www.clef-campaign.org/

Avaliação do desempenho dos sistemas participantes no CLEF'2004 em relação ao português
Escrita de dois artigos sobre essa organização e avaliação
Disponibilização da colecção CHAVE (http://www.linguateca.pt/CHAVE/)
Negociações com o jornal *Folha de São Paulo* para o CLEF'2005
Preparação da colecção deste jornal
Criação de tópicos para recolha de informação (RI) em colecções jornalísticas e na Web
Criação de perguntas e respostas para resposta automática a perguntas (RAP)
Tradução para português dos tópicos para RI geográfica e para RI de imagens
http://www.linguateca.pt/CLEF/

Organização do HAREM, primeira avaliação conjunta de reconhecimento de entidades mencionadas para o português

Discussão em conjunto da forma de proceder
Criação de uma colecção dourada para comparar com o resultado dos sistemas participantes
Criação de um conjunto de textos para avaliação, a colecção HAREM
Criação de um conjunto de métricas
Desenvolvimento de um conjunto de programas para a avaliação automática
http://www.linguateca.pt/HAREM/

Continuação do desenvolvimento dos projectos associados a recursos e informação

Corpógrafo (pólo do Porto), http://www.linguateca.pt/Corpografo/
TrAva (pólo do Porto), http://www.linguateca.pt/TrAva/
COMPARA (pólo de Oslo) http://www.linguateca.pt/COMPARA/
NATools (pólo de Braga) http://www.linguateca.pt/NATools/
Floresta Sintá(c)tica (pólos de Oslo e Braga) http://www.linguateca.pt/Floresta/
AC/DC (pólo de Oslo) http://www.linguateca.pt/ACDC/
Busca (pólos de Oslo e do Porto), http://acdc.linguateca.pt/busca/
Esfinge (pólo de Oslo) http://acdc.linguateca.pt/Esfinge/
WPT03 (pólo do XLDB), http://xldb.fc.ul.pt/linguateca/WPT_03.html

Colaboração com o projecto GREASE (pólo do XLDB), http://xldb.fc.ul.pt/index.php?page=GREASE

Participação do tumba no CLEF (pólo do XLDB)

Início dos seguintes projectos

SIEMÊS (pólo do Porto)
REPENTINO (pólo do Porto)
Anotação morfossintáctica do COMPARA (Susana Inácio)
ELLE (pólo do LabEL)
Chuveiro de dicionários (pólo de Braga)
Baco/Rede (pólo do Porto)
GKB (pólo do XLDB)

1.2 Formação

A Linguateca passou a dedicar mais tempo à formação avançada de pessoal especializado no processamento computacionald do português, quer através do trabalho dos bolseiros de doutoramento:

- Rachel Aires (último ano)
- Marcirio Chaves (segundo ano)
- Alberto Simões (primeiro ano)
- Nuno Seco (primeiro ano)

quer através da orientação dos bolseiros de investigação

- Débora Oliveira
- Luís Miguel
- Isabel Marcelino
- Susana Inácio
- Rosário Silva

Sobre o trabalho específico de cada bolseiro, veja-se os respectivos relatórios.

1.3 Encontros e conferências

Organização de encontros internos da Linguateca (apenas se relatam aquelas que deram origem a despesas de deslocação):

Vinda de Nuno Cardoso a Oslo, Agosto de 2004

Encontro em Oslo (Novembro 2004) das três bolseiras de linguística: Débora Oliveira, Isabel Marcelino e Susana Inácio

Vinda de Marcirio Chaves (doutorando) a Oslo, Abril de 2005

Simpósio doutoral em Lisboa (Maio de 2005): Marcirio Chaves, Luis Sarmento, Alberto Simões, Cristina Mota, Diana Santos

Participação em encontros e conferências (apenas se relatam aqueles cujas deslocações foram pagas pela Linguateca ou onde se apresentou material associado ao trabalho da Linguateca):

LREC2004 (Lisboa, Maio de 2004): Alberto Simões, Luís Sarmento, Luis Costa, Diana Santos, Rachel Aires, Nuno Cardoso, Anabela Barreiro, Cristina Mota, Marcirio Chaves

SIGIR 2004 (Sheffield, Inglaterra, de 25 a 29 de julho de 2004): Rachel Aires

Encontro da SEPLN (Barcelona, Agosto de 2004): Alberto Simões

CLEF 2004 (Bath, Setembro 2004): Nuno Cardoso, Paulo Rocha, Luís Costa, Diana Santos

Translation studies (Lisboa, Setembro de 2004) Diana Santos

ISLA (Lisboa, Outubro de 2004)

IBERAMIA (México, Novembro de 2004) Luís Sarmento

XATA 2005 (Braga, Fevereiro de 2005) Alberto Simões, Rui Vilela

META (Montréal, Abril 2005) Belinda Maia

2. Factores de implantação

2.1 Utilização dos serviços

Como actualização do relatório do ano anterior, apresentamos aqui de novo o tamanho do portal dedicado ao processamento computacional da língua portuguesa e o seu uso (à data de escrita deste secção do relatório, 11 de Maio de 2005).

Na tabela 1 dá-se uma panorâmica da dimensão da nossa presença na rede, discriminando o número de páginas ou documentos. Convém indicar que 1126 destes itens correspondem a documentação própria, ou seja, criada no âmbito da Linguateca.

Tabela 1: Distribuição do conteúdo do sítio, por tipo de ficheiro

Tipo	**Quantidade**
html	1122
txt	188
ps	56
pdf	43
doc	50
outros	44
Total	**1466**

Tabela 2: Ligações externas, por localização geográfica

Sufixo geográfico	Quantidade
pt	791
br	783
com	336
org	131
edu	111
de	86
uk	75
dk	50
net	47
es	45
fr	33
gov	30
no	26
it	24
ca	23

outros	121
Total	**2736**

Outra forma de avaliar a extensão do nosso serviço à comunidade é a discriminação do número de pedidos dos nossos recursos, ou seja, contar quantos investigadores ou grupos de investigação encomendaram (e receberam) os corpora ou colecções que são tornados acessíveis em texto completo:

CETEMPúblico : 324
CETENFolha : 171
CHAVE : 15
WBR-99 : 12
WPT-03 : 8

Quanto aos utilizadores registados do Corpógrafo, totalizam neste momento 415.

Quanto aos nossos serviços disponíveis na rede, que não exigem registo prévio, tiveram até agora um pouco mais de 150 mil pedidos. Veja-se http://acdc.linguateca.pt/estatisticas.html para mais informação sobre as visitas ao nosso projecto.

Finalmente, para dar uma ideia da nossa procura internacional, apresentamos na Figura 3 a distribuição do perfil geográfico dos visitantes das nossas páginas.

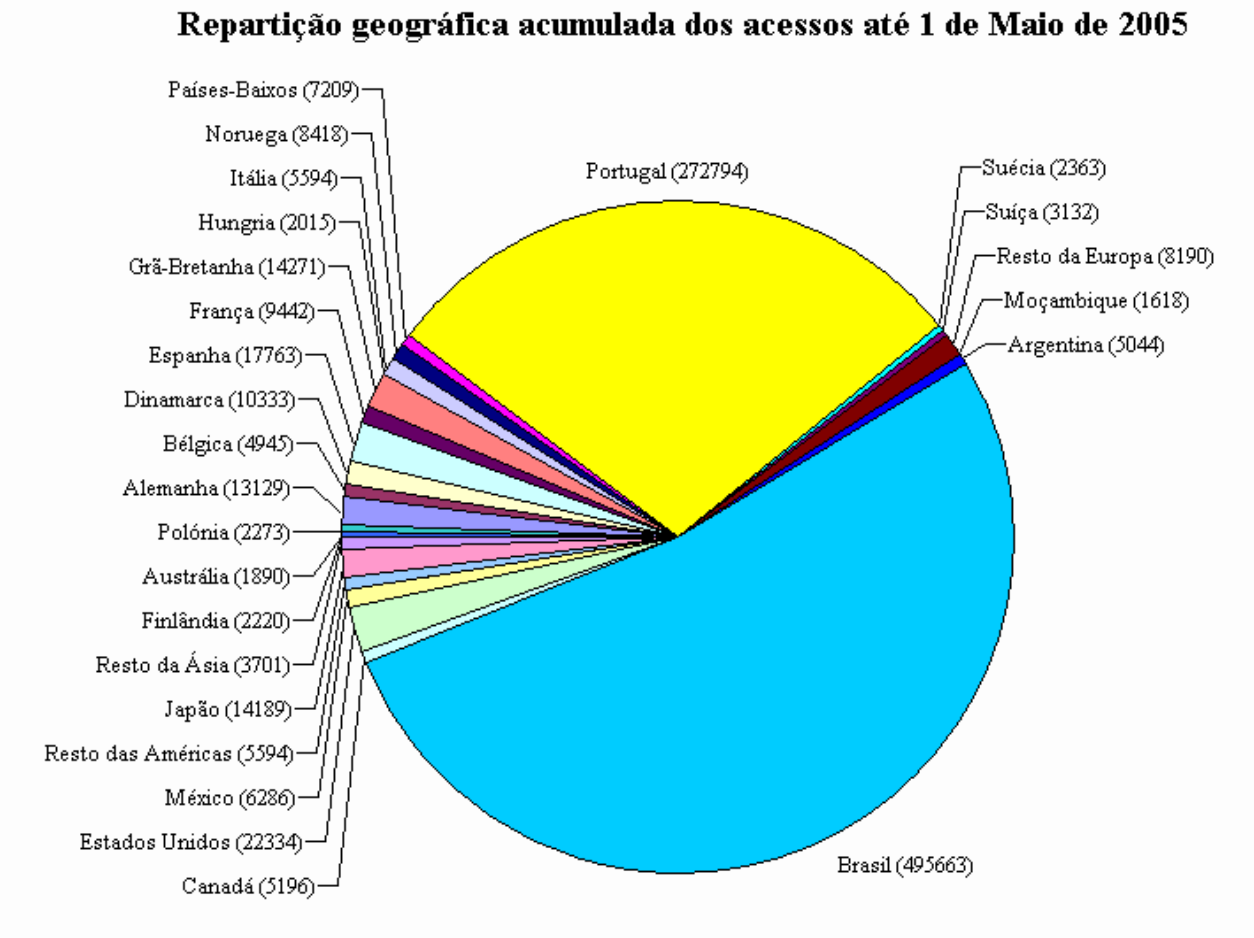

Figura 3: A distribuição geográfica dos acessos que permitem conhecer a origem (cerca de 41% do número total de acessos)

2.2 Publicações

Nesta rubrica englobamos todas as publicações que foram total ou parcialmente efectuadas no âmbito do projecto.

Publicadas no período a que se refere o presente relatório

1. Diana Santos, Belinda Maia & Luís Sarmento. "Gathering empirical data to evaluate MT from English to Portuguese". In Lambros Kranias, Nicoletta Calzolari, Gregor Thurmair, Yorick Wilks, Eduard Hovy, Gudrun Magnusdottir, Anna Samiotou & Khalid Choukri (eds.), *Proceedings of the LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora* (Lisboa, Portugal, 25 May 2004), pp. 14-17.
2. Rachel Aires, Aline Manfrin, Sandra Maria Aluísio & Diana Santos. "What Is My Style? Stylistic features in Portuguese web pages according to IR users' needs". In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of LREC 2004* (Lisboa, Portugal, 26-28 May 2004), pp. 1943-1946.
3. Diana Santos & Anabela Barreiro. "On the problems of creating a consensual golden standard of inflected forms in Portuguese". In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of LREC 2004* (Lisboa, Portugal, 26-28 May 2004), pp. 483-486.
4. Luís Sarmento, Belinda Maia & Diana Santos. "The Corpógrafo - a Web-based environment for corpora research". In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of LREC 2004* (Lisboa, Portugal, 26-28 May 2004), pp. 449-452.
5. Diana Santos. "Aonde vamos em relação a *aonde*". *The ESPecialist* **25.1** (2004). São Paulo.
6. Frankenberg-Garcia, Ana. “Lost in parallel concordances”, in Guy Aston, Silvia Bernardini and Dominic Stewart (eds.), *Corpora and Language Learners*, John Benjamins, 2004.
7. Mário J. Silva, Bruno Martins, Marcirio Silveira Chaves, Nuno Cardoso & Ana Paula Afonso. "Adding Geographic Scopes to Web Resources". In *ACM SIGIR Workshop on Geographic Information Retrieval* (Sheffield - UK, June 2004), s/ pp.
8. Alberto Manuel Simões. “Parallel Corpora word alignment and applications”. Tese de Mestrado, Universidade do Minho, 14 de Junho de 2004.
9. Alberto Manuel Simões, Xavier Gomez Guinovart & José João Almeida. "Distributed Translation Memories implementation using WebServices". In *Sociedade Espanola para el Procesamiento del Lenguaje Natural (SEPLN)* (Barcelona, Julho, 2004).
10. Nuno Cardoso, Mário J. Silva & Miguel Costa. "The XLDB Group at CLEF 2004". In Carol Peters & Francesca Borri (eds.), *Cross Language Evaluation Forum: Working Notes for the CLEF 2004 Workshop (CLEF 2004)* (Bath, UK, 15-17 September), Pisa, Italy: IST-CNR, pp. 183-191.
11. Alessandro Vallin, Bernardo Magnini, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Simov & Richard Sutcliffe. "Overview of the CLEF 2004 Multilingual Question answering track". In Carol Peters & Francesca Borri (eds.), *Cross Language Evaluation Forum: Working Notes for the CLEF 2004 Workshop (CLEF 2004)* (Bath, UK, 15-17 September), Pisa, Italy: IST-CNR, pp. 281-294.

12. Luís Costa. "First evaluation of Esfinge - a question-answering system for Portuguese". In Carol Peters & Francesca Borri (eds.), *Cross Language Evaluation Forum: Working Notes for the CLEF 2004 Workshop (CLEF 2004)* (Bath, UK, 15-17 September), Pisa, Italy: IST-CNR, pp. 393-402.
13. Diana Santos & Paulo Rocha. "CHAVE: topics and questions on the Portuguese participation in CLEF". In Carol Peters & Francesca Borri. *Cross Language Evaluation Forum: Working Notes for the CLEF 2004 Workshop* (15-17 September, Bath, UK), IST-CNR, Pisa, Italy, pp.639-48.
14. Rachel Aires. "O uso de características lingüísticas para a apresentação dos resultados de busca na Web de acordo com a intenção da busca do usuário", *In IX Simpósio de teses e dissertações do ICMC-USP São Carlos* (São Carlos (SP), 19 e 20 de Novembro de 2004).
15. Diana Santos, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela & Susana Afonso. "Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa", in Guillermo De Ita Luna, Olac Fuentes Chávez, Mauricio Osorio Galindo (eds.), *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", IX Iberoamerican Conference on Artificial Intelligence (IBERAMIA), November 2004, Puebla, Mexico*, pp. 147-154.
16. Alberto Manuel Simões, Tiago Bezerra & Pedro Henriques. "A importância das Ontologias num Museu Virtual". In *II Congresso Internacional de Investigação e Desenvolvimento Sócio-cultural* (Paredes de Coura, Outubro de 2004), CD-ROM, s/pp.
17. Rui Vilela, Alberto Manuel Simões, Eckhard Bick & José João Almeida. "Representação em XML da Floresta Sintáctica". In José Carlos Ramalho, Alberto Simões & João Correia Lopes (eds.), *3ª Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas (XATA 2005)* (Braga, Fevereiro de 2005), Departamento de Informática, Universidade do Minho, pp. 351-361.

Completadas e enviadas para publicação, mas ainda não publicadas

1. Diana Santos. "Breves explorações num mar de língua", *Ilha do Desterro*, no prelo.
2. Diana Santos, Eckhard Bick, Raquel Marchi & Susana Afonso. "A Floresta Sintá(c)tica como recurso para estudar a língua portuguesa", in Regina Gerber & Vera Vasilévski (orgs.), *Um percurso para pesquisas com base em corpus*. Florianópolis/SC, Brasil: Editora da UFSC, no prelo.
3. Belinda Maia, Luís Sarmento & Diana Santos. "O Corpógrafo", *Terminómetro*, no prelo.
4. Luís Costa & Diana Santos. "A Linguateca e o projecto Processamento Computacional do português", *Terminómetro*, no prelo.
5. Diana Santos & Paulo Rocha. "The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE". In Carol Peters, Paul Clough, Julio Gonzalo, Gareth Jones, Michael Kluck & Bernardo Magnini (eds.), *Advances in Cross-Language Information Retrieval: Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)* (Bath,

UK, 15-17 September 2004), Heidelberga, Alemanha: Springer. Lecture Notes in Computer Science. Revised version of Santos & Rocha (2004)

6. Nuno Cardoso, Mário J. Silva & Miguel Costa. "The XLDB Group at CLEF 2004". In Carol Peters, Paul Clough, Julio Gonzalo, Gareth Jones, Michael Kluck & Bernardo Magnini (eds.), *Advances in Cross-Language Information Retrieval: Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)* (Bath, UK, 15-17 September 2004), Heidelberga, Alemanha: Springer. Lecture Notes in Computer Science. Revised version of Cardoso et al. (2004)
7. Luís Costa. "First Evaluation of Esfinge - a Question Answering System for Portuguese". In Carol Peters, Paul Clough, Julio Gonzalo, Gareth Jones, Michael Kluck & Bernardo Magnini (eds.), *Advances in Cross-Language Information Retrieval: Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)* (Bath, UK, 15-17 September 2004), Heidelberga, Alemanha: Springer. Lecture Notes in Computer Science. Revised version of Costa (2004)
8. Alessandro Vallin, Bernardo Magnini, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Simov & Richard Sutcliffe. "Overview of the CLEF 2004 Multilingual Question answering track". In Carol Peters, Paul Clough, Julio Gonzalo, Gareth Jones, Michael Kluck & Bernardo Magnini (eds.), *Advances in Cross-Language Information Retrieval: Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)* (Bath, UK, 15-17 September 2004), Heidelberga, Alemanha: Springer. Lecture Notes in Computer Science. Revised version of Vallin et al. (2004)
9. Belinda Maia & Luís Sarmento. "The Corpógrafo - an Experiment in Designing a Research and Study Environment for Comparable Corpora Compilation and Terminology Extraction". In *Proceedings of eCoLoRe / MeLLANGE Workshop, Resources and Tools for e-Learning in Translation and Localisation* (Centre for Translation Studies, University of Leeds, UK, 21-23 March 2005).
10. Ana Frankenberg-Garcia. "Pedagogical uses of Monolingual and Parallel Concordances". *English Language Teaching Journal* **59.3**.
11. Diana Santos. "Introdução ao modelo de avaliação conjunta". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.
12. Luís Costa, Paulo Rocha & Diana Santos. "Organização e resultados morfolímpicos". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.
13. Elisabete Ranchhod & Cristina Mota. "Unidades lexicais multipalavra, um osso duro de roer: Sobre a participação do LabEL". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.
14. José João Almeida & Alberto Simões. "Jspellando nas morfolimpíadas: Sobre a participação do Jspell ". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.
15. Viviane Moreira Orengo & Diana Santos. "Radicalizadores versus analisadores morfológicos: Sobre a participação do Removedor de Sufixos da Língua Portuguesa ". In

Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.

16. Anabela Barreiro & Susana Afonso. "Construção da lista dourada para as primeiras Morfolimpíadas do português". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.
17. Luís Sarmento, Anabela Barreiro, Belinda Maia & Diana Santos. "Avaliação de Tradução Automática: alguns conceitos e reflexões". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.
18. Luís Sarmento. "Ferramentas para experimentação, recolha e avaliação de exemplos de tradução automática ". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.
19. Belinda Maia & Anabela Barreiro. "Uma experiência de recolha de exemplos classificados de tradução automática de inglês para português". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.
20. Alberto Simões & José João Almeida. "Avaliação de alinhadores à frase ". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.
21. Rachel Virgínia Xavier Aires & Sandra Maria Aluísio. "As avaliações atuais de sistemas de busca na Web e a importância do usuário". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005
22. Cristina Mota, Diana Santos & Elisabete Ranchhod. "Avaliação de reconhecimento de entidades mencionadas: princípio de AREM ". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.
23. Paulo Rocha & Diana Santos. "CLEF: Abrindo a porta à participação internacional em RI do português". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.
24. Nuno Cardoso, Bruno Martins, Daniel Gomes & Mário J. Silva. "WPT03: Recolha da Web portuguesa". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.
25. Eckhard Bick, Diana Santos, Susana Afonso & Rachel Marchi. "Floresta Sintá(c)tica: Realidade ou ficção? ". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. 2005.

Completadas e enviadas para apreciação

26. Rachel Aires, Sandra Aluísio & Diana Santos. “Web pages classification according to users’ intention”, em apreciação.
27. Diana Santos & Nuno Cardoso. “Assumptions and evaluation resources of HAREM – Evaluation Contest for Named Entity Recognizers for Portuguese”, em apreciação.
28. Marcírio Silveira Chaves, Mário J. Silva & Bruno Martins. "A Geographic Knowledge Base for Text Processing", em apreciação.

29. Rachel Aires. “Bringing the user into the search engine internals”, em apreciação.
30. Luís Sarmento & Ana Pinto. "Acquiring definition patterns and lists of secondary terms for the automatic extraction of definitions in Portuguese", em apreciação.
31. Luís Sarmento, Ana Sofia Pinto, Luís Cabral & Débora Oliveira. "REPENTINO - A collaborative gazetteer for Named Entity Recognition in Portuguese", em apreciação.
32. Rachel Aires & Sandra Aluísio. “Leva-e-traz: A user-aware meta-search engine”, em apreciação.

Aceites para publicação, ainda em redacção à data deste relatório

33. Belinda Maia. "Terminology and Translation -Bringing Research and Professional Training Together Through Technology". In *META Simposium - For a Proactive Translatology* (Université de Montréal, Québec, Canadá, 7-9 April 2005).
34. Luís Sarmento. "A Simple and Robust Algorithm for Extracting Terminology". In *META Simposium - For a Proactive Translatology* (Université de Montréal, Québec, Canadá, 7-9 April 2005).
35. Rachel Aires, Sandra Aluísio & Diana Santos. ""Yes, user!": compiling a corpus according to what the user wants". In *Corpus Linguistics 2005* (Birmingham, 14-17 July 2005).
36. Débora Oliveira, Luís Sarmento, Belinda Maia & Diana Santos. "Corpus analysis for indexing: when corpus-based terminology makes a difference". In *Corpus Linguistics 2005* (Birmingham, 14-17 July 2005).
37. Ana Frankenberg-Garcia. "A corpus-based study of loan words in original and translated texts". In *Corpus Linguistics 2005* (Birmingham, 14-17 July 2005).

Relatórios on-line

1. Susana Afonso. "Árvores deitadas: Descrição do formato e das opções de análise na Floresta Sintáctica". 2004. http://www.linguateca.pt/Floresta/ArvoresDeitadas.pdf
2. Luís Sarmento. "Pré-Especificação de Requisitos do Sucessor do TrAva". Porto, 10 de Setembro de 2004. http://poloclup.linguateca.pt/docs/6f.pdf.
3. Susana Afonso. "A Floresta Sintá(c)tica como recurso". 2004. http://www.linguateca.pt/Floresta/Afonso2004Recurso.doc.
4. Ana Frankenberg-Garcia. "COMPARA English Tutorial". 28 September 2004. http://www.linguateca.pt/COMPARA/Tutorial.doc.
5. Eckhard Bick. "Looking at the Floresta Sintá(c)tica with a CorpusEye: A user-friendly cross-language search interface". 2004. http://www.linguateca.pt/Floresta/floresta-corpuseye_en.doc.
6. Raquel Marchi. "Revisão humana da Floresta Sintá(c)tica: exemplos e método". 2004. http://www.linguateca.pt/Floresta/Marchi2004Revisao.doc.

7. Rachel Aires, Aline Manfrin, Sandra Aluísio & Diana Santos. “Which classification algorithm works best with stylistic features of Portuguese in order to classify web texts according to users’ needs?” Relatório técnico nº 241, outubro de 2004, ICMC/USP, http://www.nilc.icmc.usp.br/nilc/download/airesetaltr0409.pdf.
8. Débora Oliveira & Ana Sofia Pinto. Extracção de definições no Corpógrafo. Outubro de 2004. http://poloclup.linguateca.pt/~doliveira/publico/corpografo.doc.
9. Ana Frankenberg-Garcia. "COMPARA - Aula Prática em Português". 14 de Março de 2005. http://www.linguateca.pt/COMPARA/AulaPratica.doc.
10. Luís Sarmento. Descrição técnica do REPENTINO. Relatório Técnico Linguateca (em preparação), Maio 2005, http://poloclup.linguateca.pt/docs/rel_repentino.doc.
11. Marcírio Chaves, Bruno Martins & Mário J. Silva. "Grease Knowledge Base". DI/FCUL TR 05—XX : Departmento de Informática, Universidade de Lisboa. A ser disponibilizado em Junho de 2005.

Apresentações em encontros ou conferências que não deram origem a publicação

1. Belinda Maia, "Trabalhando com o CORPÓGRAFO: no ensino e na pesquisa", apresentado no *IV Encontro de Corpora - Da compilação à anotação lingüística, padronização e análise de corpora*, Universidade de São Paulo, Brasil, 23 de Agosto, 2004. http://www.nilc.icmc.usp.br/ivencontro/slide_belinda.ppt.
2. Ana Frankenberg-Garcia. "O corpus COMPARA". *IX Encontro Nacional / III Encontro Internacional de Tradutores* (Fortaleza (CE), Setembro 2004). http://www.linguateca.pt/Repositorio/FrankenbergGarcia2004-Fortaleza.ppt.
3. Diana Santos. "Working with Portuguese corpora". (Universitet i Oslo, 22 October 2004). Extended version of a presentation at ISLA, in Lisbon 1 October 2004. http://www.linguateca.pt/Diana/download/KRI2004.pdf.
4. Alberto Simões. “EBMT: Example Based Machine Translation”. Simpósio Doutoral do Departamento de Informática da Universidade do Minho, Dezembro de 2004, http://alfarrabio.di.uminho.pt/~albie/publications/ebmt-poster.ps.gz.
5. Rachel Aires. ”Apresentação sobre a criação do corpus de necessidades”, USP, 2 de maio de 2005.

Apresentações em meios de comunicação social

1. Alberto Simões, “A Linguateca em Braga”, *UM Jornal*, ano 2, n. 7, 5 de Agosto de 2004, Universidade do Minho.

3. Relatório do pólo de Oslo

Luís Costa e Diana Santos
Outros elementos da equipa afectos ao pólo: Paulo Rocha, Rachel Aires, Susana Afonso, Raquel Marchi, Susana Inácio e Rosário Silva.

O pólo de Oslo tem a responsabilidade da orientação do projecto, manutenção do portal e de diversos serviços e recursos associados ao mesmo. Tem também estado envolvido em diversas iniciativas de avaliação de recursos relacionados com o processamento computacional do português. Segue-se uma descrição mais pormenorizada de cada uma destas vertentes:

3.1 Orientação do projecto

Foram regularmente discutidos, com os responsáveis de cada pólo, quais as directrizes a seguir no trabalho a desenvolver, bem como a avaliação do trabalho já feito. Uma questão que tem tido sempre uma particular atenção é a tentativa de que os pólos trabalhem em conjunto, ou seja que aproveitando a especialização de cada um deles em diferentes áreas do processamento do português, o trabalho produzido em cada um deles possa ser integrado no trabalho de cada um dos outros.

3.2 Manutenção do portal

O portal tem vindo a ser constantemente enriquecido com nova informação.

À data da escrita deste relatório, o catálogo de recursos contém 792 entradas, o catálogo de ferramentas tem 125 entradas, o catálogo de actores tem 426 entradas, o catálogo de publicações tem 1009 entradas e existem 149 entradas relativas a informação interessante.

Continuamos a manter um fórum com informação sobre conferências, notícias, cursos e bolsa de emprego na área do processamento computacional do português. Mantemos também um arquivo das entradas mais antigas (342 peças até agora).

Algum esforço tem também sido dispendido na actualização dos bastidores, ou seja, a documentação técnica associada à gestão do portal.

Finalmente, algum trabalho foi dedicado à melhoria da documentação dos recursos no repositório da Linguateca, em particular associado à história do Corpus NILC, no qual a maior parte dos recursosactuais para o português brasileiro têm origem.

3.2.1 Busca

O sistema de busca e as páginas associadas foram aperfeiçoados, adicionando a funcionalidade de busca geral (por texto livre), a possibilidade – ainda experimental – de buscas por palavras-chave (em colaboração com o pólo do Porto, ver relatório de Débora Oliveira).

Foram também criadas versões em inglês de todas as páginas do sistema de busca.

Finalmente, o sistema de busca de publicações foi extensivamente melhorado, permitindo vários modos de apresentação dos resultados, e um maior rigor na criação das listas de publicações associadas à Linguateca.

3.3 Desenvolvimento e manutenção de serviços e recursos

3.3.1 Projecto AC/DC

Além da habitual resposta a utilizadores e ajuda na procura, foram calculadas e disponibilizadas listas de frequências de palavras para o WPT-03 (colecção da rede portuguesa) e WBR-99 (colecção da rede brasileira) (criadas pelo Nuno Seco, do pólo de Lisboa).

Foi calculada também a frequência de formas e lemas por categoria gramatical nos corpora AC/DC, e melhorado o acesso directo ao projecto.

3.3.2 Projectos CETEMPúblico e CETENFolha

Foram disponibilizados mediante um registo prévio os primeiros 300.000 extractos do corpus CETEMPúblico anotado pelo analisador sintáctico PALAVRAS de Eckhard Bick, além da habitual disponibilização do CETENFolha anotado e de ambos os corpora sem anotação.

A anotação sintáctica da totalidade do CETEMPúblico, após instalação do anotador sintáctico PALAVRAS, de Eckhard Bick, em Oslo, encontra-se em progresso.

3.3.3 Projecto COMPARA/DISPARA

O corpus COMPARA, em colaboração com Ana Frankenberg-Garcia e a sua equipa de estagiárias, foi constantemente sendo melhorado e aumentado, tendo no ano a que se refere este relatório sido disponibilizadas 3 novas versões.

De momento o corpus está na versão 6.1, com 58 pares de textos, e mais de um milhão de palavras em cada

Além disso, foi iniciada a anotação sintáctica do COMPARA (parte portuguesa) pela Susana Inácio (ver relatório desta bolseira), orientada principalmente pelo pólo de Oslo, que desenvolveu além disso uma nova interface, assim como especificou a forma de prosseguir nesse trabalho, como relatar problemas e qual a periodicidade da actualização do corpus.

3.3.4 Projecto Floresta Sintá(c)tica

O projecto Floresta Sintá(c)tica também continuou a ser desenvolvido, tendo visto a instalação de mais cinc novas versões este ano.

De momento está na versão 6.6, com 8.258 árvores, correspondentes a quase 200 mil palavras revistas, assim como a Floresta Virgem (um milhão de palavras em poruguês de Portugal e outro tanto em português do Brasil) foi também disponibilizado publicamente.

Entre as várias tarefas, saliente-se que o texto das frases referente a cada árvore foi revisto e reposto, o documento principal de documentação foi tornado acessível na rede, e várias melhorias ao sistema Águia foram implementadas, no que se refere a mensagens de erro e a uma melhor visualização das concordâncias.

Além disso, o pólo de Braga (ver relatório respectivo) iniciou a tarefa de validação da Floresta Sintáctica, além de a passar a disponibilizar em formatos XML.

3.3.5 Esfinge

Adaptação do sistema de resposta automática a perguntas Esfinge para a participação no CLEF'2004:

- Foi necessário codificar as colecções de documentos fornecidas pela organização e o sistema foi adaptado para indicar também um código de documento da colecção para justificar as respostas, o que era um dos requisitos para participar nesta avaliação conjunta.
- Envio de dois conjuntos de resultados para as 199 perguntas criadas pela organização.

- Análise dos resultados obtidos, com a consequente escrita de um artigo, de um poster, e da reformulação do primeiro artigo para publicação internacional

Após o estudo do desempenho do esfinge no CLEF 2004, acrescentaram-se novos padrões para tratar perguntas que não foram tratadas correctamente no CLEF'2004 e corrigiu-se a forma como o analisador morfológico Jspell estava a ser invocado.

Iniciou-se a preparação da participação no CLEF'2005, de forma a que neste ano para além da tarefa pt-pt (perguntas em português, respostas em português), seja também possível participar nas tarefas en-pt e pt-en (perguntas em inglês, respostas em português e vice versa).

3.4 Actividades associadas a avaliação conjunta

3.4.1 HAREM

O HAREM, primeira avaliação de sistemas de reconhecimento de entidades mencionadas para português foi organizada em estreita colaboração com o pólo do XLDB (ver relatório do pólo)

As tarefas em que o pólo de Oslo participou significativamente foram

- Escolha e contabilização primária da maior parte dos textos da colecção dourada e da colecção HAREM.
- Especificação e documentação dos critérios de definição de uma EM,
- Revisão manual da colecção dourada e discussão dos casos complicados
- Especificação e documentação das métricas
- Documentação da arquitectura e estrutura dos programas de avaliação
- Comunicação com os participantes

3.4.2 Organização do livro sobre avaliação conjunta

Todo o trabalho de chamada de artigos, informação aos autores, organização da revisão cruzada, revisão tipográfica e bibliográfica, e edição electrónica foi feito no pólo de Oslo, levando a um conjunto de 24 capítulos, mais prefácio e bibliografia, que já foi enviado para uma editora.

3.4.3 Disponibilização da colecção CHAVE

A colecção CHAVE foi um dos resultados da participação da Linguateca na organização do CLEF 2004.

Além dos textos completos do PÚBLICO de 1994 e 1995, contém uma lista de cinquenta tópicos em português, compilados em cooperação com os restantes organizadores do CLEF; as avaliações (binárias) de cada tópico; uma lista de 700 perguntas e respostas em português, compiladas em cooperação com os restantes organizadores do QA@CLEF; um conjunto não-exaustivo de documentos que suporta a(s) resposta(s) para um subconjunto de 199 dessas perguntas.

A actividade de documentação assim como de disponibilização da própria colecção às pessoas que se registam passou pois a fazer parte das tarefas de rotina

4. Relatório do pólo de Braga

Rui Vilela e Alberto Simões
Responsável por parte do DI: José João Dias de Almeida.

O pólo de Braga está sediado no Departamento de Informática da Universidade do Minho, integrado na actividade de processamento de linguagem natural desta instituição de ensino, no âmbito do projecto Natura. Rui Vilela substituiu Alberto Simões como colaborador da Linguateca a tempo inteiro no pólo em Setembro de 2004.

O pólo tem vários objectivos, de entre os quais se destacam:

- Desenvolvimento de ferramentas para os recursos existentes da Linguateca.
- Distribuição e desenvolvimento de aplicações públicas.
- Desenvolvimento de programas para validar a qualidade dos recursos existentes.

Uma actividade com algum peso corresponde ao estudo e desenvolvimento de ferramentas associadas a corpora bilingues, associado ao mestrado (concluído a 14 de Junho de 2004 e arguido em Setembro) e doutoramento em curso do Alberto Simões, este último descrito no relatório correspondente.

4.1 Desenvolvimento e disponibilização de ferramentas

O pólo de Braga no ano a que se refere o relatório, criou ou aumentou os seguintes recursos:

4.1.1 NATools

O NATools é um pacote de ferramentas para o alinhamento de corpora paralelos ao nível da frase e da palavra. É baseado no trabalho de Djoerd Hiemstra no Twente Aligner.

Foram implementadas as seguintes melhorias/novidades:

- Alinhamento eficiente com divisão de corpora em fatias para alinhamento de corpora de grandes dimensões;
- Consulta on-line de dicionários probabilísticos;
- Consulta on-line de corpora paralelos com realce de paralelismo entre palavras;
- Inclusão do alinhador à palvra Vanilla Aligner de Gale e Church;
- Integração de alguns filtros para pré-processamento de corpora;
- Criação de uma shell para interagir com o sistema de alinhamento;
- Criação de índices inversos para pesquisa eficiente de exemplos de tradução (a usar brevemente em várias ferramentas já desenvolvidas)

4.1.2 Chuveiro de dicionários

O chuveiro de dicionários é um projectoque foi iniciado em Janeiro de 2005 para disponibilizar versões actualizadas do dicionário português para correcção ortográfica em formatos de código aberto (open source).

São gerados periodicamente versões dos seguintes dicionários: MySpell (usado no OpenOffice, Mozilla, Firefox, Thunderbird), Ispell, Aspell, Jspell, que são disponibilizados na página da Linguateca do pólo de Braga.

4.1.3 Memórias de tradução distribuídas

As memórias de tradução distribuídas pretendem ser um serviço na rede prestado quer por empresas de tradução, comunidades de tradutores ou mesmo tradutores independentes em que cada tradutor possa, através da rede, consultar as memórias de outros tradutores.

O projecto Memórias de Tradução Distribuídas pretende desenvolver uma base de implementação de servidores e clientes para a troca Cliente/Servidor de unidades de tradução. Em relação a este projecto foram publicados dois artigos e realizados vários testes, nomeadamente em:

- criação de um protótipo em Perl e usando o indexador Glimpse;
- indexação de exemplos em pequenos pacotes para rapidez de resposta (será brevemente substituído pelo NATools agora que suporta índices inversos de pesquisa);
- manutenção de um servidor em Braga;
- colaboração com a Universidade de Vigo para a implementação de um sistema de MTD usando PHP e Glimpse.

4.2 Avaliação e controlo de qualidade

No âmbito das actividades de avaliação conjunta da Linguateca, o pólo de Braga participou na organização do HAREM, asim como se constitui em validador independente da Floresta Sintáctica.

4.2.1 Validação da Floresta Sintáctica

O pólo de Braga desenvolveu ferramentas para validação da floresta e exportação para outros formatos (Penn Treebank, Tiger-XML, SQL).

Além disso, aplica-as sempre que novas versões são instaladas, produzindo relatórios de erros que são enviados para a equipa de revisão.

4.2.2 HAREM

Apoio ao pólo XLDB no desenvolvimento das aplicações para avaliação no HAREM.

4.3 Desenvolvimento de recursos

4.3.1 Projecto AC/DC: novos corpora

Produção de dois novos corpora:

- Museu da Pessoa: Conjunto de 109 entrevistas do Museu da Pessoa Português disponibilizados pelo Núcleo Português do Museu da Pessoa, http://www.museudapessoa.net
- Condiv: Conjunto de 3982 artigos jornalísticos de diversos jornais desportivos portugueses e brasileiros (A Bola, Mundo Desportivo, O Jogo, Record, Jornal de Sports, Estado de São Paulo, Gazeta, O Lance) dos anos 50, 70 e 2000 baseado num estudo contrastivo do léxico. Trabalho elaborado por Augusto Soares da Silva, ao abrigo de um projecto de investigação liderado por este. O pólo de Braga deu algum apoio informático, e construiu o corpus no formato AC/DC (ainda se encontram pendentes algumas autorizações legais, contudo).

5. Relatório do pólo do Porto

Luís Sarmento
Outros elementos da equipa afectos ao pólo: Débora Oliveira, Luís Miguel Cabral, Ana Sofia Pinto, Anabela Barreiro.
Responsável por parte da FLUP: Belinda Maia.

À data da elaboração deste relatório, o pólo do Porto encontra-se já a meio do seu terceiro ano de existência e tem vindo a desenvolver actividades em vários domínios do processamento computacional do português. O pólo conta actualmente com três colaboradores a tempo-inteiro suportados directamente pelo projecto: Luís Sarmento, Luís Cabral (bolseiro de informática desde Julho de 2004), Débora Oliveira (bolseira de linguística desde Julho de 2004), e conta também com a colaboração em tempo parcial da bolseira Ana Sofia Pinto, que, apesar de ser suportada pela Faculdade de Letras, está totalmente integrada nas actividades do Pólo. Tem também interagido com o Pólo a estudante de doutoramento Anabela Barreiro, co-orientada por Belinda Maia, e cujo o trabalho de investigação se integra nos temas de desenvolvimento abordados no Pólo, nomeadamente, a avaliação de tradução automática.

A partir de Março de 2003, o pólo tem vindo a concentrar os seus esforços em dois projectos principais: o Corpógrafo e o desenvolvimento de estruturas para a Avaliação de Tradução Automática (ATA). Com a entrada da bolseira Débora Oliveira deu-se também início a uma nova linha de desenvolvimento com o objectivo de melhorar o sistema Busca, projecto que tem várias interacções com o projecto Corpógrafo.

Ultimamente, desde o início de 2005, e com o crescente aumento de competências e capacidade produtiva da equipa, o Pólo do Porto tem feito alguma investigação aplicada a áreas mais específicas, para além dos projectos de desenvolvimento que já se encontravam sob a sua alçada. Neste domínio destacam-se o desenvolvimento de um sistema experimental para reconhecimento de entidades nomeadas, que participou no HAREM, e o desenho e desenvolvimento de bases de dados lexicais geradas a partir de corpora da Linguateca (CETEMPúblico e WPT03) e que permitem apoiar estudos e experiências que envolvam pesquisas massiças em corpora.

5.1 O Corpógrafo

O Corpógrafo entrou agora no seu terceiro ano de desenvolvimento e possui perspectivas de alargamento fora do seu território inicial que tem sido a FLUP. A versão actual é a V2. Encontram-se inscritos neste momento 405 utilizadores, embora só cerca de metade é que possam ser considerados utilizadores habituais. Nos últimos meses, o número de utilizadores habituais tem subido significativamente com inscrições provenientes da FLUP, da FLUL, da USP e de várias outras universidades do Brasil.

Está neste momento a ser terminado o porte de toda a infraestrutura de informação para um sistema de base de dados standard que permite uma maior estabilidade e um melhor desempenho do Corpógrafo e que resultará na sua versão 3.

O porte está a envolver a reprogramação de grande parte das camadas de comunicação entre o armazém de dados e a interface com o utilizador pelo que se trata de uma tarefa demorada, embora fundamental nesta altura. Em paralelo, aproveitou-se para reformular o esquema de dados do Corpógrafo, melhorando algumas lacunas que foram apontadas pelos utilizadores, em particular a nível da gestão terminológica. A interface com o utilizador tem

também sido alvo de muitas melhorias e a documentação de apoio ao utilizador está igualmente a ser reformulada para acompanhar todas estas alterações. Espera-se lançar a V3 até ao final de Maio.

5.1.1 O SAGI

O SAGI tem vindo a ser implementado desde Junho de 2004 sendo a principal tarefa do bolseiro Luís Miguel Cabral. O desenvolvimento da maioria das funcionalidades do SAGI foi terminado no final do ano de 2004 e desde então têm sido feitas pequenas melhorias que resultam de necessidades impostas pela sua utilização no Corpógrafo. Mais informação acerca do SAGI será fornecida no relatório do bolseiro Luís Cabral.

5.1.2 Melhoramentos no Busca

Desde Julho de 2004, tem sido desenvolvido trabalho para a melhoria do motor de pesquisa Busca, integrado no sítio web da Linguateca. A bolseira Débora Oliveira foi contratada para desenvolver os estudos linguísticos associados ao objectivo de encontrar formas mais eficientes de proceder à indexação, ordenação de resultados e processamento de expressões de pesquisa. O trabalho tem sido realizado em colaboração com o pólo de Oslo que assegura a implementação técnica das soluções propostas e a integração dos recursos criados no pólo do Porto.

Para mais detalhes sobre este tópico remetemos o leitor para o relatório da bolseira Débora Oliveira.

5.2 METRA, TrAva, CorTA e Boomerangue

Durante o período a que se refere este relatório não se fez praticamente nenhum desenvolvimento substancial relativamente a estas ferramentas e recursos, exceptuando algumas tarefas essenciais de manutenção e algumas reformulações simples, e muita documentação (ver abaixo). A causa do abrandamento da actividade nesta área está relacionada com outras prioridades que entretanto foram surgindo.

No entanto, a este nível o pólo não está totalmente parado, pois foi produzido um estudo sobre os possíveis desenvolvimentos a partir do estado actual, e onde se traçam alguns cenários futuros, estudo este disponível através do sítio web do pólo. Além disso, têm sido mantidas reuniões com alguns alunos de mestrado da Prof. Belinda Maia que estão a realizar estudos no âmbito da avaliação de tradução automática e na caracterização de padrões de utilização de serviços de tradução automática.

Espera-se poder re-arrancar estes projectos em breve depois da finalização da Versão 3 do Corpógrafo.

No período a que se refere o presente relatório, um esforço considerável foi dedicado à produção e revisão cruzada de vários capítulos do livro sobre o Avalon nesta área, produzindo os seguintes capítulos, relativos a questões associadas à avaliação da tradução automática:

- Luís Sarmento, Anabela Barreiro, Belinda Maia e Diana Santos. Avaliação de tradução automática: alguns conceitos e reflexões;
- Luís Sarmento. Ferramentas para experimentação, recolha e avaliação de exemplos de tradução automática;
- Belinda Maia e Anabela Barreiro. Uma experiência de recolha de exemplos classificados de tradução automática de inglês para português.

5.3 Organização e identificação de entidades lexicais

5.3.1 O SIEMÊS

O pólo do Porto participou na actividade de Avaliação Conjunta organizada pela Linguateca, o HAREM, levando à avaliação o seu sistema de reconhecimento de entidades mencionadas, o SIEMÊS: Sistema de Identificação de Entidades Mencionadas com Estratégia Siamesa. O SIEMÊS foi inicialmente desenvolvido usando as ideias associadas ao extractor de terminologia usado no Corpógrafo, embora tenha depois sido desenvolvido usando ideias que resultaram das discussões mantidas durante o período preparatório do HAREM. O SIEMÊS, como o nome indica, tenta identificar e reconhecer entidades segundo uma estratégia dupla: (i) procurando semelhanças usando uma vasta base de dados de exemplos de entidades nomeadas, já categorizados segundo a sua semântica intrínseca (i.e. ignorando totalmente o contexto da ocorrência); e (ii) desambiguando as possibilidades obtidas no passo anterior, usando regras de análise de contexto.

O SIEMÊS foi já disponibilzado internamente para uso por outros pólos, nomeadamente a participação do pólo de Oslo no QA@CLEF com o sistema Esfinge, e a sua disponibilização pública encontra-se iminente. Para mais informação sobre o SIEMÊS, por favor consultar o seu sítio de rede, onde se encontra uma demonstração do sistema e também explicações técnicas acerca do seu funcionamento: http://poloclup.linguateca.pt/siemes/.

5.3.2 O REPENTINO

O REPENTINO, REPositório para reconhecimento de ENTIdades NOmeadas é um recurso que resulta dos desenvolvimentos realizados no Pólo do Porto no âmbito da preparação para o HAREM. Durante o processo de discussão de quais as categorias de entidades a considerar e de quais as entidades que se incluiriam nessas categorias foi feita uma recolha interna de exemplos de entidades. Com o objectivo de reutilizar ao máximo esse esforço, decidiu-se organizar os exemplos recolhidos numa base de dados que incluiria exemplos de entidades das várias categorias e subcategorias consideradas relevantes e que teria um acesso público via web.

Neste momento, o REPENTINO armazena cerca de 450 mil exemplos organizados em 11 categorias e 100 subcategorias. O REPENTINO encontra-se acessível publicamente para consulta e para obtenção do recurso completo em formato XML, através do seu sítio oficial na rede em: http://poloclup.linguateca.pt/repentino/.

5.3.3 BACO - Base de Co-ocorrências

O BACO é um sistema vocacionado para a pesquisa flexível de co-ocorrências significativas em texto, sendo que estes dados podem ser úteis em várias aplicações. O BACO pretende explorar a enorme oferta dos corpora da Linguateca e aproveitá-los para estes estudos de co-ocorrências. O cerne do BACO é uma enorme base de dados que armazena vários tipos de co-ocorrências em texto e permite pesquisar esses dados de uma forma rápida: é obtida velocidade de pesquisa em troca de uma elevada redundância no esquema dos dados armazenados.

O projecto BACO está ainda a ser documentado, tendo sido pela primeira vez apresentado à restante equipa da Linguateca no seminário doutoral em Lisboa a 8 de Maio.

5.4 Contactos Externos

Desde 2003, têm vindo também a ser desenvolvidos contactos com instituições e parceiros exteriores à FLUP. Os contactos tem sido maioritariamente liderados pela Prof. Belinda Maia, com o apoio do restante pessoal do Pólo, e têm permitido não só aumentar a visibilidade do

projecto, como também lançar bases para futuras parcerias. Para além da divulgação do projecto Linguateca em geral e dos seus recursos, o foco principal destes contactos tem sido a promoção da utilização do Corpógrafo por outras instituições académicas e de investigação. A este nível destacam-se apresentações e colaborações com instituições de ensino como a Universidade de São Paulo (Setembro de 2004) e Universidade de Tampere (Abril de 2005) e apresentações em workshops e forums de tradução como III Encontro Internacional de Tradutores em Fortaleza (Agosto de 2004) e como o Workshop eColore (Leeds, Março de 2005).

Foi também preparada em parceria com o INEGI uma candidatura a financiamento de projecto pela FCT. O projecto tinha como objectivo de usar as capacidades de construção de bases de dados terminológicas oferecidas pelo Corpógrafo como base para o desenvolvimento de repositórios documentação especializada pesquisáveis de uma forma inteligente. Apesar do projecto não ter sido aprovado, os contactos com o INEGI têm continuado no sentido de se preparar futuras colaborações.

Adicionalmente, e após vários contactos com responsáveis da Universidade Pompeu Fabra de Barcelona, encontra-se também em preparação uma eventual instalação na referida instituição de uma versão do Corpógrafo. Aguarda-se, contudo, a clarificação de determinados detalhes burocráticos para poder avançar.

A nível da Universidade do Porto, foram também realizados bastantes contactos na tentativa de aumentar a base de apoio do pólo e de poder vir a aumentar a equipa. Os principais contactos foram realizados com a Reitoria da Universidade do Porto, onde apresentamos o projecto e tentamos promover eventuais interacções com outros grupos. Como resultado desta apresentação, o Pólo do Porto foi convidado a participar num encontro patrocionado pela Reitoria onde se discutiu as possibilidades da criação de um centro de ciências cognitivas no seio da universidades. Aguardamos ainda desenvolvimentos da discussão gerada a partir desta iniciativa.

A nível interno na FLUP, têm sido realizadas diversas apresentações, quer a alunos, quer a docentes da FLUP, com o objectivo de estimular a utilização das ferramentas desenvolvidas e estimular o interesse para a área do processamento computacional do português. A este nível sentimos que neste terceiro ano o trabalho começa a dar alguns frutos com um aumento significativo do interesse por questões de linguística computacional por parte dos alunos e dos docentes. Para além do Corpógrafo, outras ferramentas da Linguateca (AC/DC, COMPARA, METRA) têm vindo ser regularmente utilizadas por docentes e alunos no decorrer das correspondentes actividades curriculares.

6. Relatório do pólo de Lisboa no XLDB

Nuno Cardoso
Outros elementos da equipa afectos ao pólo: Marcírio Chaves, Nuno Seco.
Responsável por parte do XLDB: Mário J. Silva.

O pólo de Lisboa no XLDB tem como principais objectivos organizar, documentar e manter as colecções para avaliação de recuperação de informação existentes para o português. As actividades principais prendem-se com:

- Estabelecer rotinas de medição e auscultação da Web em relação à língua portuguesa
- Desenvolver programas que aumentem a integração entre as actividades da Linguateca e do Grupo XLDB, em particular acoplando funções linguísticas ao Tumba!, o motor de procura do XLDB
- Co-organizar o HAREM- Avaliação conjunta para o Reconhecimento de Entidades Mencionadas
- Ajudar à implantação do paradigma de avaliação conjunta na RI do português

6.1 Actividades de avaliação conjunta

6.1.1 HAREM

HAREM - Co-organização do **HAREM** - **A**valiação (conjunta de sistemas) de **R**econhecimento de **E**ntidades **M**encionadas, em conjunto com os pólos de Oslo e de Braga da Linguateca. Envolveu:

1. Selecção e extracção de textos das colecções Web (WPT 03 e WBR 99)
2. Compilação dos textos, categorização e formatação em SGML
3. Etiquetação de textos da Colecção Dourada. Envio e recepção de textos etiquetados pelos participantes, e a revisão dos mesmos.
4. Criação e manutenção de um sítio na rede, para a divulgação de notícias, calendários, motivações, directivas, resultados e outros doumentos, para os participantes e para a comunidade científica interessada nesta avaliação conjunta.
5. Divulgação do evento, recepção de pedidos de participação e *feedback.*
6. Elaboração de diversas directivas, para discussão conjunta, sobre as regras de etiquetagem, a definição de categorias, as tarefas a implementar, as regras de avaliação, as metodologias a adoptar na avaliação, formatos das colecções, entre outros.
7. Criação de protótipos de avaliação de resultados e de validadores de submissões, para a criação de relatórios de resultados dos participantes.
8. Documentação e disponibilização da Colecção Dourada, uma colecção para a avaliação de sistemas de reconhecimento de entidades mencionadas.

Além disso, algum apoio foi dado (por elementos distintos dos da comissão de organização do HAREM, bem entendido) à participação do XLDB no HAREM, através da utilização do GKB, parcialmente desenvolvido pela Linguateca, no sistema CAGE.

6.1.2 CLEF

Participação do XLDB nesta avaliação conjunta (a primeira envolvendo RI para português) em 2004, na tarefa "Adhoc Monolingue PT". Esta participação envolveu:

1. Carregamento e indexação das colecções do CLEF
2. Modificação de módulos do tumba! para a criação de resultados para o CLEF

3. Envio e análise de resultados
4. Escrita de um artigo sobre a participação do XLDB no CLEF2004.

Participação (em curso) no evento de 2005, nas tarefas: Adhoc Monolingue PT, Bilingue EN-PT, GeoCLEF e WebCLEF. Esta participação envolveu, até ao momento:

2. Carregamento e Indexação das colecções
3. novas aproximações, incorporando o que se aprendeu em 2004, usando novos recursos e programas de processamento de linguagem natural (jspell e baco), e envolvendo o pólo de Braga da Linguateca na sub-tarefa de tradução de textos, bem como alterações a componentes do tumba!, resultantes da análise dos resultados no CLEF 2004 e TREC 2004.
4. Participação na criação de tópicos em Português, para a comunidade CLEF.

6.1.3 TREC

Participação do XLDB no TREC de 2004, na tarefa WebTrack. Esta participação envolveu:

1. Desenvolvimento de uma proxy, para avaliar o desempenho dos batedores (crawlers) do tumba!.
2. Carregamento das colecções
3. Indexação das colecções e avaliação do processo de indexação e ordenação do tumba!.
4. Criação de resultados, recepção e análise de resultados, avaliação destes e detecção de pontos fracos a serem melhorados no tumba!.

6.2 Desenvolvimento e disponibilização de recursos

WPT 03

2. Criação da WPT 03 a partir das recolhas feitas pelo tumba!.
3. Tratamento e disponibilização de uma lista de registos de acesso ao servidor web do tumba!, contendo termos de pesquisa dos utilizadores do tumba! durante 6 meses
4. Criação de programas de auxílio à manipulação da colecção
5. Divulgação da WPT 03 pelos meios de comunicação social
6. Distribuição da WPT 03 a várias instituições, com fins de investigação
7. Actividade de obtenção de feedback por parte dos utilizadores, com vista ao melhoramento da colecção e ferramentas
8. Cálculo de frequências de palavras e de outras estatísticas da WPT 03
9. Trabalho em comum com o pólo do Porto para interrogar a WPT03 em MySQL
10. Planeamento de uma nova colecção da Web Portuguesa

6.3 Participação no projecto GREASE

Uma das sinergias entre a Linguateca e o XLDB é o projecto GREASE, no âmbito do qual um dos doutoramentos se inclui. Nesse projecto a contribuição do pólo da Linguateca foi a seguinte:

1. Desenvolvimento da base de conhecimento geográfico (GKB - Grease Knowledge Database), para ser usada por sistemas de processamento de textos e para apoiar a detecção do âmbito de uma página Web.
2. Produção de documentação extensa sobre o GKB.
3. Expansão da base de conhecimento para suporte multilingue.
4. Geração de algumas ontologias geográficas com base nesta
5. Algumas tentativas de detecção de entidades geográficas no WPT 03, com base no conhecimento contido no GKB.

7. Relatórios dos bolseiros

7.1 Rachel Virgínia Xavier Aires, bolseira de doutoramento

A bolsa refere-se ao meu projecto de doutorado, "Linguarudo: O uso de características lingüísticas para a apresentação dos resultados de busca na Web de acordo com a intenção de busca do usuário – uma instanciação para o português", orientado por Diana Santos e Sandra Aluísio, que vem sido apoiado pela FCCN desde setembro de 2001.

7.1.1 Criação de recursos e rodagem de experimentos

- Acompanhamento e testes da ferramenta de apoio para construção de coleções de teste construída pelo aluno de iniciação científica da Universidade de São Paulo, Mateus Godoi Milanez.
- Aumento do corpus de treinamento. O corpus foi inicialmente aumentado para 1760 páginas. Revisão da classificação das 1760 páginas do corpus de treinamento classificado por necessidades de usuários, encontrar (i) definições ou explicações sobre como e/ou porque algo acontece; (ii) explicações sobre como fazer algo ou como algo é feito; (iii) panorama sobre um determinado assunto; (iv) notícias; (v) informações sobre pessoas ou organizações e (vi) serviços prestados online. O objetivo desta revisão foi verificar se não existiam páginas: que de alguma forma fugissem do padrão de classificação estabelecido; páginas repetidas; páginas de diferentes variantes da língua portuguesa.
- O corpus de necessidades foi reclassificado de acordo com as classes: (i) ultraformal, formal, semiformal ou informal; (ii) contextualizado ou não e (iii) opinião ou descrição.
- Experimentos com todas as 1252 páginas do corpus fornecido por José Martins Junior (compilado para a sua tese: "Classificação de páginas na internet", Universidade de São Paulo, abril de 2003)
  - utilizando nosso conjunto de 46 features e os algoritmos J48, SMO e LMT do Weka. Os testes foram realizados utilizando *10 folds cross-validation*.
  - utilizando as 11205 *features* utilizadas por ele em sua dissertação
- Experimentos com o corpus Lácio-ref, (i) dividido por gêneros, (ii) dividido por tipos textuais. Para estes experimentos foi necessário deividir o corpus em arquivos separados e não em um único arquivo. Foram utilizados os algoritmos J48, SMO e LMT.
- Criação de três corpora, um que distingue os textos como ultraformais, formais, semiformais ou informais, outro que distingue como contextualizados ou não e, o último que faz distinção entre textos que emitem opinião e textos que são somente descritivos.
- Criação de um corpus com textos relacionados à área de direito de dois tipos: (i) textos voltados para pessoas que trabalham com direito, como juízes e advogados, por exemplo, pareceres e medidas provisórias; (ii) textos voltados para pessoas comuns, que falam sobre direito, mas não são técnicos, como notícias em revistas e manuais para o consumidor sobre seus direitos. O corpus foi composto por 200 textos.
- Criação de um corpus de textos de direito em inglês, também composto por 200 páginas, construído para ilustrar em um artigo a abordagem que estamos seguindo neste trabalho.
- Construção de um script para cálculo de 60 *features* estilísticas de textos em inglês.
- Experimentos com corpora de direito em inglês e português utilizando os algoritmos J48, SMO e LMT.

7.1.2 Criação do protótipo de um meta-buscador

- Estudo de código aberto em Java disponibilizado por outros projetos e sistemas: Lucene[1], Nutch[2], Carrot2[3], Objectssearch[4] e Egothor[5]. Lucene é uma biblioteca de código aberto para busca. Nutch é aplicação deste código aberto para a busca na Web que utiliza Lucene, porém não é um site de busca, e pode ser utilizado tanto para a Web como para Intranets. Carrot2 é um *framework* para clusterização, que inclui além dos componentes para esta tarefa, um componente para meta-busca. Objectsearch é uma máquina de busca de código aberto que utiliza Nutch, Lucene, Carrot2, e outras bibliotecas de código aberto. Egothor, também é uma máquina de busca de código aberto, que pode ser configurada como uma máquina *standalone*, como um meta-buscador, como um HUB *peer-to-peer* ou como uma biblioteca para outras aplicações que precisem de busca em textos.
- Iniciamos a construção do protótipo utilizando componentes do Carrot2.
- Desenvolvimento de um questionário para investigar os tipos de classificação usados nos experimentos e quais os mais intuitivos para um usuário
    - Aplicação de questionário a: alunos de graduação de computação, de medicina, de letras e de especialização em fotografia. (As perguntas são as mesmas para todos os alunos, a diferença entre os questionários está apenas nos exemplos que ilustram algumas das perguntas.)
    - Análise dos questionários aplicados.
- Design da interface do protótipo.
- Desenvolvimento do protótipo de meta-buscador. Todas as fases da busca pelo meta-buscador foram desenvolvidas: (i) a submissão das consultas ao Google e AlltheWeb, (ii) a recuperação dos resultados, (iii) a transformação de arquivos para somente texto que reúne os scripts desenvolvidos e a (iv) a interpretação das regras de classificação.
- Atualmente estamos integrando todas as fases para gerar o protótipo.

7.1.3 Produção escrita

Rachel Aires, Aline Manfrin, Sandra Maria Aluísio & Diana Santos. "What Is My Style? Stylistic features in Portuguese web pages according to IR users' needs". In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of LREC 2004* (Lisboa, Portugal, 26-28 May 2004), pp. 1943-1946.

Rachel Aires. "O uso de características lingüísticas para a apresentação dos resultados de busca na Web de acordo com a intenção da busca do usuário". In *IX Simpósio de teses e dissertações do ICMC-USP São Carlos* (São Carlos (SP), 19 e 20 de Novembro de 2004).

Rachel Aires, Aline Manfrin, Sandra Aluísio & Diana Santos. "Which classification algorithm works best with stylistic features of Portuguese in order to classify web texts according to users' needs?" Relatório técnico nº 241, outubro de 2004, ICMC/USP, http://www.nilc.icmc.usp.br/nilc/download/airesetaltr0409.pdf.

Diana Santos, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias

[1] http://jakarta.apache.org/lucene/
[2] http://www.nutch.org
[3] http://sourceforge.net/projects/carrot2/
[4] http://www.objectssearch.com/en/about.html
[5] http://www.egothor.org/

de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela & Susana Afonso. "Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa", in Guillermo De Ita Luna, Olac Fuentes Chávez, Mauricio Osorio Galindo (eds.), *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", IX Iberoamerican Conference on Artificial Intelligence (IBERAMIA), November 2004, Puebla, Mexico*, pp. 147-154.

Rachel Virgínia Xavier Aires & Sandra Maria Aluísio. "As avaliações atuais de sistemas de busca na Web e a importância do usuário". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*.

Rachel Aires, Sandra Aluísio & Diana Santos. ""Yes, user!": compiling a corpus according to what the user wants". In *Corpus Linguistics 2005* (Birmingham, 14-17 July 2005).

Rachel Aires, Sandra Aluísio & Diana Santos. “Web pages classification according to users’ intention”, em apreciação.

Rachel Aires. “Bringing the user into the search engine internals”, em apreciação.

Rachel Aires & Sandra Aluísio. “Leva-e-traz: A user-aware meta-search engine”, em apreciação.

Rachel Aires. Apresentação sobre a criação do corpus de necessidades na USP em 2 de maio de 2005, para os alunos de pós-graduação da disciplina Tópicos de Inteligência Artificial.

7.2 Marcirio Silveira Chaves, bolseiro de doutoramento

Meu trabalho de doutoramento está no âmbito do projeto GREASE, o qual visa atribuir âmbitos geográficos a sites na web. Neste primeiro ano como bolseiro desenvolvi uma base de conhecimento geográfico, a qual é utilizada por sistemas de processamento de texto. Além disso, realizei diversos experimentos no sentido de encontrar padrões geográficos em texto na língua portuguesa utilizando os nomes armazenados na base de conhecimento construída.

Dentre os diversos aspectos positivos do trabalho, destaco meu amadurecimento no desenvolvimento de trabalhos científicos (escrita de artigos, participação em eventos, etc...). A integração do trabalho da tese em um projeto de pesquisa que envolve um grupo de pesquisadores também é um ponto que merece destaque e que tem me permitido avançar com meu trabalho. Além disso, o trabalho que está sendo realizado em algumas disciplinas poderá ser incorporado na tese.

Por outro lado, eu esperava ter cursado todas as disciplinas no segundo semestre de 2004. Por motivo de disponibilidade das mesmas não foi possível. Mas neste semestre pretendo cumprir os créditos exigidos pelo programa de pós-graduação.

Apesar de o trabalho estar evoluindo dentro do cronograma inicial estabelecido, eu esperava ter avançado mais na revisão bibliográfica até o momento, já tendo produzido um quantidade maior de material escrito.

A seguir descrevo o trabalho realizado:

7.2.1 Participação ativa no projeto GREASE

Como já foi indicado, a minha tese está integrada no projeto GREASE, que pretende melhorar e dar maiores funcionalidades ao tumba! Nessa ótica, relato as atividades práticas relacionadas com o projeto:

- desenvolvimento de uma base de conhecimento geográfico (GKB – Grease Knowledge Base) que está sendo usada por sistemas de processamento de texto;
- expansão da base de conhecimento para suportar termos multi-língua;
- geração de ontologias para outros sistemas do projeto utilizarem;
- documentação da GKB, a ser publicada inicialmente como relatório técnico.

7.2.2 Atividades relacionadas com a escrita da tese

- revisão de bibliografia (em progresso)
- três apresentações realizadas em Oslo, sobre ontologias, extração de ontologias a partir textos em linguagem natural, e sobre o sistema KnowItAll; e melhoradas no simpósio doutoral em Lisboa
- desenho de alguns experimentos para medir o problema da detecção de entidades geográficas no WPT03, antes e durante o período de formação em Oslo
- início da implementação dos experimentos desenhados

7.2.3 Participação no HAREM – Avaliação de Reconhecimento de Entidades Mencionadas

Extração de uma ontologia a partir da GKB. Essa ontologia foi utilizada como uma fonte de dados para auxiliar o sistema CAGE (CApturing Geographical Entities) na tarefa de reconhecimento de entidades mencionadas.

Esta é uma primeira forma de ajuizar e avaliar o uso da GKB (e da ontologia gerada) na solução de problemas práticos.

Embora a participação no HAREM não fosse diretamente relacionada com o meu trabalho, faz parte integrante do GREASE e da forma como se pode aproveitar a sinergia entre as actividades da Linguateca e do XLDB.

7.2.4 Disciplinas cursadas no âmbito do doutoramento

Tópicos Avançados em Sistemas de Informação I – Aprovado (17 valores)
Essa disciplina me permitiu fazer uma panorâmica sobre o estado da arte na área de Sistemas de Informação. Trabalhei os seguintes assuntos: métodos de classificação automática, usabilidade em grupo, anotação semântica na web semântica, palestras na web, fotos digitais. O assunto de anotação semântica na web semântica está mais próximo do tema que estou investigando na tese.
Grafos (em andamento)
A abordagem adotada na disciplina está relacionada a área de Pesquisa Operacional. Tenho estudado diversos tipos de grafos, bem como seus comportamentos. A base de conhecimento (GKB) que estou construindo para a tese é exportada para outros sistemas em forma de grafo. Assim, as estratégias para percorrimento de grafos estudadas na disciplina podem ser aplicadas ao grafo gerado a partir da GKB.
Introdução à Investigação em Inteligência Artificial (em andamento)
Estou trabalhando a parte de introdução de senso comum a sistemas computacionais. Está área está em um estágio preliminar em termos de sistemas sendo utilizados na prática na língua inglesa. Para o português, ainda não tenho conhecimento de nenhuma iniciativa. Nesse sentido, pretendo explorar o uso da base de conhecimento construída por agentes automáticos.
Linguística Computacional II (em andamento)
O conteúdo desta disciplina fornece suporte à construção de dicionários eletrônicos e gramáticas com ênfase na língua portuguesa. Tenho construído pequenas gramáticas para descoberta de padrões geográficos e reconhecimento de entidades mencionadas geográficas.

7.2.5 Publicações

Mário J. Silva, Bruno Martins, Marcirio Silveira Chaves, Nuno Cardoso & Ana Paula Afonso. "Adding Geographic Scopes to Web Resources". In *ACM SIGIR Workshop on Geographic Information Retrieval* (Sheffield - UK, June 2004), s/ pp.

Diana Santos, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela & Susana Afonso. "Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa", in Guillermo De Ita Luna, Olac Fuentes Chávez, Mauricio Osorio Galindo (eds.), *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", IX Iberoamerican Conference on Artificial Intelligence (IBERAMIA), November 2004, Puebla, Mexico*, pp. 147-154.

Marcírio Chaves, Bruno Martins & Mário J. Silva. "Grease Knowledge Base". DI/FCUL TR 05—XX : Department of Informatics, University of Lisbon. Maio de 2005.

Marcírio Silveira Chaves, Mário J. Silva & Bruno Martins. "A Geographic Knowledge Base for Text Processing". Enviado para apreciação.

7.2.6 Participação em atividades de formação

Conferência LREC2004 em Lisboa, Maio de 2004.

Semana em Oslo, Abril de 2005.
Simpósio doutoral em Lisboa, Maio de 2005.

7.3 Alberto Simões, bolseiro de doutoramento

Este relatório refere-se ao trabalho realizado no doutoramento de Alberto Simões sob a orientação conjunta de José João Almeida e Diana Santos, sob o tema de "Tradução Baseada em Exemplos", iniciado em Setembro de 2004.

7.3.1 Trabalho realizado

- Obtenção da licença do Part-Of-Speech tagger TnT[6], que irá ser usado durante o projecto de doutoramento;
- Melhoria do conjunto de ferramentas NATools, bem como incorporação de nova funcionalidade para facilitar testes de alinhamento com corpora filtrados de diferentes formas. Incorporação de índices inversos para pesquisa eficiente de exemplos. (O NATools (http://natura.di.uminho.pt/natura/natura/NATools) é um pacote de ferramentas desenvolvidas durante a tese de mestrado e que irá ser melhorado e enriquecido ao longo do doutoramento. De momento, inclui um alinhador de corpora ao nível de frase, e um alinhador ao nível da palavra.)
- Realização de alinhamentos ao nível da palavra a diversos corpora paralelos (nomeadamente ao EuroParl e COMPARA) para extracção de dicionários probabilísticos de tradução a serem usados no sistema de tradução a implementar.
- Aplicação de diferentes metodologias de alinhamento de corpora paralelos para a obtenção de dicionários probabilísticos com ênfase em diferentes categorias de palavras e expressões terminológicas.
- Colaboração com o pólo do XLDB na participação no CLEF, com uma ferramenta de tradução.

7.3.2 Publicações relacionadas com o trabalho da tese

Alberto Simões. "EBMT: Example Based Machine Translation". Poster para apresentação no Simpósio Doutoral do Departamento de Informática da Universidade do Minho, Dezembro de 2004, http://alfarrabio.di.uminho.pt/~albie/publications/ebmt-poster.ps.gz.

Alberto Simões & José João Almeida. "Enhancing Word Alignment with simple NLP tools" enviado à *SEPLN 2005, XXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural*.

José João Almeida & Alberto Simões. "Bootstrapping a Multilingual Dictionary using Thesaurus tools" enviado à *SEPLN 2005, XXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural*.

Alberto Simões, "TABE, tradução automática baseada em exemplos", apresentação no simpósio doutoral em Lisboa a 5 de Maio de 2005.

7.3.3 Outras publicações

Alberto Manuel Simões, Xavier Gomez Guinovart & José João Almeida. "Distributed Translation Memories implementation using WebServices". In *Sociedade Espanola para el Procesamiento del Lenguaje Natural (SEPLN)* (Barcelona, Julho, 2004), pp. 89-94.

[6] http://www.coli.uni-sb.de/~thorsten/tnt/

José João Almeida & Alberto Simões. "Jspellando nas morfolimpíadas: Sobre a participação do Jspell ". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa.*

Alberto Simões & José João Almeida. "Avaliação de alinhadores à frase ". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*.

Alberto Manuel Simões. “Parallel Corpora word alignment and applications”. Tese de Mestrado, Universidade do Minho, 14 de Junho de 2004.

Alberto Manuel Simões, Tiago Bezerra & Pedro Henriques. "A importância das Ontologias num Museu Virtual". In *II Congresso Internacional de Investigação e Desenvolvimento Sócio-cultural* (Paredes de Coura, Outubro de 2004), CD-ROM, s/pp.

Diana Santos, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela & Susana Afonso. "Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa", in Guillermo De Ita Luna, Olac Fuentes Chávez, Mauricio Osorio Galindo (eds.), *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", IX Iberoamerican Conference on Artificial Intelligence (IBERAMIA), November 2004, Puebla, Mexico*, pp. 147-154.

Rui Vilela, Alberto Manuel Simões, Eckhard Bick & José João Almeida. "Representação em XML da Floresta Sintáctica". In José Carlos Ramalho, Alberto Simões & João Correia Lopes (eds.), *3ª Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas (XATA 2005)* (Braga, Fevereiro de 2005), Departamento de Informática, Universidade do Minho, pp. 351-361.

José João Almeida & Alberto Simões. “Inferindo tipos a partir de documentos XML”. In José Carlos Ramalho, Alberto Simões & João Correia Lopes (eds.), *3ª Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas (XATA 2005)* (Braga, Fevereiro de 2005), Departamento de Informática, Universidade do Minho, pp.144-154.

7.4 Nuno Alexandre Lopes Seco, bolseiro de investigação/doutoramento

Neste relatório faz-se um breve resumo do trabalho desenvolvido pelo bolseiro desde 15 de Dezembro de 2004 até ao presente.

Antes de especificar o trabalho realizado convém esclarecer que o mesmo tem sido efectuado com uma bolsa de Técnico de Investigação. Esta situação deve-se ao facto de o grau de Mestre em Ciências da Computação ainda não ter sido atribuído pela University College Dublin, Irlanda. A tese já foi arguida e existem pequenas rectificações a realizar antes da publicação final. Note-se que estas rectificações em nada põem em causa a atribuição do grau, cuja certificação se espera ser obtida dentro de poucos meses.

O trabalho desenvolvido ao longo deste período pode ser categorizado segundo duas perspectivas diferentes:

1. Trabalho preliminar para o doutoramento.
2. Trabalho relacionado com a integração na Linguateca.

Ambas as dimensões são igualmente importantes. O trabalho desenvolvido com a presente bolsa visa ser continuado no âmbito de uma bolsa de doutoramento, daí julga-se importante tentar definir o mais possível o rumo a seguir no posterior doutoramento. O segundo ponto também é importante pois permite ao bolseiro familiarizar-se com a Linguateca e com os recursos disponibilizados. Esta familiarização, para além de contribuir directamente para Linguateca, permite ao bolseiro identificar recursos que lhe poderão ser úteis durante o doutoramento e/ou a falta de outros que será preciso desenvolver.

7.4.1 Trabalho preliminar para o doutoramento

Em relação a este ponto o bolseiro começou por se debruçar sob o tema de Categorização de Textos (“clustering”). O bolseiro efectuou uma revisão bibliográfica preliminar da área de modo a familiarizar-se com o estado-da-arte.

De momento participa no desenvolvimento de um categorizador de texto capaz de identificar se um texto está escrito em português de Portugal ou do Brasil.

Dada a criação do WPT03 pelo pólo de Lisboa no XLDB, que disponibilizou um conjunto de procuras feitas ao tumba! pelos utilizadores, o bolseiro também iniciou uma análise desses diários, de modo a tentar obter informação sobre os interesses dos utilizadores.

Por outro lado, ao familiarizar-se com os recursos existentes para o processamento computacional do português, identificou-se uma lacuna. Julga-se que a criação de uma ontologia lexical para o português poderá ser uma mais valia para a comunidade de investigadores que estudam a nossa língua. Aliás, recursos semelhantes já existem para outras línguas, sendo talvez a mais conhecida o WordNet para o Inglês. Existem já algumas iniciativas como o WordNet.PT (http://www.instituto-camoes.pt/bases/lingua/wordnet.htm) e o WordNet.BR (http://www.nilc.icmc.usp.br/nilc/projects/wordnetbr.htm), mas todas as tentativas de obtenção destes recursos têm sido em vão, indicando que não são públicos nem provavelmente virão a ser.

O trabalho elaborado pelo bolseiro no seu mestrado pode servir de suporte ao desenvolvimento de tal recurso linguístico. Trabalho esse que passou por analisar recursos semelhantes (mas para a língua inglesa) e utilizar um desses recursos para o desenvolvimento de um sistema computacional linguístico. Salienta-se ainda o facto de o artigo intitulado "Creative discovery in the lexical validation gap", do qual o bolseiro é co-autor, ter sido recentemente aceite para publicação no *Journal of Computer Speech and Language*. Embora este trabalho não esteja relacionado com o processamento do português, julgamos que muito do conhecimento

adquirido trabalhando com recursos deste género poderá ser reutilizado no desenvolvimento de um recurso semelhante para o português.

7.4.2 Trabalho relacionado com a integração na Linguateca

No que diz respeito ao trabalho relacionado com a integração na Linguateca, o bolseiro tem auxiliado a organização do HAREM (Avaliação conjunta de sistemas de Reconhecimento de Entidades Mencionadas).

Inicialmente começou por ajudar na selecção de textos que iriam fazer parte da Colecção Dourada (uma colecção de textos manualmente etiquetada com informação léxico-semântica).

Actualmente encontra-se a implementar alguns dos programas que têm como objectivo avaliar a participação dos vários concorrentes.

Por fim, o bolseiro implementou um conjunto de programas que permitiu calcular as frequências das palavras existentes no WPT03 e que agora é disponibilizado no sítio da Linguateca.

7.4.3 Publicações

Jer Hayes, Nuno Seco and Tony Veale. Creative discovery in the lexical validation gap. Aceite para publicação em *Journal of Computer speech and Language* (Special Issue on Multiword Expressions).

7.5 Isabel Marcelino, bolseira de investigação

Este relatório refere-se à actividade da bolseira, no âmbito da bolsa AnELL, de 1 de Outubro de 2004 a 14 de Maio de 2005.

7.5.1 Melhoramentos ao AnELL

Foram implementadas as seguintes modificações:

- Melhoramento de alguns dicionários.
- Alteração da ordem pela qual os dicionários são aplicados.
- Modificação do programa que chama os dicionários.
- Melhoramentos gerais associados a formatação e análise de expressões não incluídas nos dicionários, mas muito frequentes nos textos (por exemplos, URLs, alíneas, listas numeradas, etc.)

Foram sugeridas, além disso, as seguintes alterações, associadas ao melhoramento das interfaces do AnELL.

- Interface mais simples para cada modo.

Finalmente, foi iniciada uma definição clara das especificações de um sistema de ajuda à revisão para melhorar a eficiência do modo supervisionado.

7.5.2 Revisão de um pedido

Anotação e revisão de um pedido feito ao AnELL no modo supervisionado (pedido do Professor Tony Sardinha).

7.5.3 Desenvolvimento do ELLE a partir de um protótipo já existente

Criação do sistema ELLE (Etiquetador LaBEL-Lex de Entidades (mencionadas)), adaptando e melhorando um sistema desenvolvido pela professora Elisabete Ranchhod para a constituição da colecção dourada do HAREM.

Participação no HAREM com oELLE. (Este sistema anota entidades mencionadas referentes a pessoas, locais, acontecimentos e organizações.)

Actualização dos dicionários de nomes próprios do AnELL com as entradas encontradas no HAREM.

7.5.4 Actividades de formação

Participação no workshop «Working with language corpora», no ISLA, Lisboa, 1 de Outubro de 2004.

Formação de 15 dias pela Cristina Mota (Outubro de 2004), que incidiu sobre o AnELL, o Intex e o Linux.

Formação em Oslo em Novembro de 2004, que incidiu sobre a Linguateca e os projectos AC/DC, Floresta Sintá(c)tica, Trava, Busca, Corpógrafo, Esfinge e COMPARA, assim como clarificação do trabalho a fazer sobre o AnELL e como melhorá-lo.

7.5.5 Produção escrita

- Elaboração de um relatório com a descrição das tarefas a desenvolver
- http://www.linguateca.pt/Equipa/isabel/Relatorio_Oslo.doc
- Aperfeiçoamento e conclusão do ficheiro de ajuda ao serviço «Leia-me_Notaçõe».
- Criação da documentação do ELLE (em progresso).
- Documentação dos recursos linguísticos subjacentes ao AnELL (em progresso).

7.6 Susana Inácio, bolseira de investigação

Este relatório refere-se à actividade da bolseira, no âmbito da bolsa COMPARA, de 1 de Outubro de 2004 a 14 de Maio de 2005.

7.6.1 Trabalho realizado

O trabalho principal no âmbito da bolsa é a revisão da anotação morfossintáctica do COMPARA, ou seja, a validação da anotação criada por um analisador sintáctico (*parser*) automático, de forma a ter informação gramatical fidedigna no COMPARA:

Tal tarefa pode ser discriminada no seguintes pontos:

- Familiarização com os corpora portugueses já anotados da Linguateca, tais como o AC/DC, e a Floresta Sintá(c)tica, e com o estilo de anotação subjacente (o fornecido pelo PALAVRAS e seu pós-processamento automático no AC/DC), assim como com o *IMS Corpus Workbench*, sistema de processamento de corpora que está subjacente ao COMPARA
- Participação no desenho de uma interface inicial (sistema de procura com novas funcionalidades e listas de palavras ordenadas por frequência) para ajudar à revisão do corpus, em www.linguateca.pt/COMPARA/bastidores/RevisaoSusana.html
- Participação na definição do processo de incorporar incrementalmente a revisão no corpus anotado, que implica:
  - Envio dos ficheiros emendados, para Oslo, por FTP, de quinze em quinze dias,
  - Envio de problemas encontrados no texto do corpus não anotado, por correio electrónico
  - Envio de problemas sistemáticos encontrados na anotação automática ou no seu pós-processamento, por correio electrónico
  - Documentação dos critérios e dúvidas
  - Ir buscar novos ficheiros num processo de sincronização e/ou quando novos pares são adicionados ao corpus

Como resultado, foi feito, apenas para o lado português, a revisão das seguintes formas em contexto. Para dar uma estimativa do trabalho realizado e da dimensão do que falta, apresento na tabela seguinte, a dimensão do corpus a rever e o que já foi revisto:

Cat. gram.	Número de palavras distintas	Número total de palavras	Número de palavras distintas revistas	Número total de palavras revistas
N	13872	219949	8957	214973
A	4728	59485	-	-
V	12785	228126	-	-

Estes números são indicativos, visto que já houve várias versões do PALAVRAS que poderão mudar um pouco o número de cada categoria gramatical identificada (correcta ou incorrectamente).

7.6.2 Formação

- Participação no workshop *Working with Language Corpora*, no ISLA, Lisboa, no dia 1 de Outubro de 2004.
- Formação de 3 dias, monitorizada pela Cristina Mota, no LabEL, em Outubro de 2004, sobre o AnELL e o *Intex*.
- Formação em Oslo (Novembro de 2004), sobre a Linguateca e os projectos AC/DC, Floresta Sintá(c)tica, Trava, Busca, Corpógrafo, Esfinge e COMPARA, tendo, também, como finalidade, proporcionar-me um contacto mais estreito com as tarefas a

desempenhar, ajudando-me, simultaneamente, a definir um método de trabalho para ser posto em prática na revisão da anotação morfossintáctica do COMPARA.

7.6.3 Produção escrita

Elaboração de um relatório do 1º trimestre, como bolseira.
Elaboração de um relatório de visita a Oslo, disponível em:
http://www.linguateca.pt/Equipa/SusanaInacio/RelatorioVisitaOslo.doc
Início da criação de um manual de Ajuda do COMPARA anotado, para o utilizador e para futuros anotadores.
Documentação de apoio à revisão da anotação morfossintáctica do COMPARA, em progresso, contendo de momento os seguintes capítulos:

- Distinção entre nomes e adjectivos
- Distinção entre numerais e adjectivos
- Identificação e delimitação de nomes próprios

7.7 Luís Cabral, bolseiro de investigação

Este relatório refere-se à actividade do bolseiro, no âmbito da bolsa Corpógrafo, de 3 de Julho a 14 de Maio de 2005.

7.7.1 Desenvolvimento do SAGI - Sistema de Apoio à Gestão de Interfaces

O SAGI é um projecto de raiz para facilitar a criação e gestão das interfaces de aplicações Web. Este desenvolvimento do SAGI decorreu em duas fases distintas.

Após a fase de especificação, deu-se início à implementação gradual da plataforma e das suas funcionalidades fulcrais, e à criação de uma interface de administração.

Na segunda fase, já com retorno dos utilizadores, procedeu-se à melhoria das interfaces e correcção de conflitos com outras aplicações (firewalls e anti-vírus), continuando o desenvolvimento de métodos que facilitem a reestruturação do Corpógrafo.

7.7.2 Corpógrafo

Assistência ao na reestruturação do Corpógrafo. Nomeadamente:

- Especificação na estrutura das bases de dados de administração e individuais.
- Criação da API para acesso aos dados.
- Reestruturação das interfaces.
- Reestruturação do código de alguns scripts.

7.7.3 TrAva – Reavaliação de resultados

Criação de programas e interfaces Web para permitir a linguistas (num primeiro momento, tal foi testado pelas bolseiras da Linguateca) avaliar uma lista de frases que continham problemas de tradução e que tinham sido previamente recolhidas usando o TrAva.

7.7.4 REPENTINO - Repositório para o Reconhecimento de Entidades Nomeadas

O REPENTINO é um sistema que armazena exemplos de entidades nomeadas, agrupando-as segundo uma taxonomia pré-definida.

- Criação de scripts de análise, inserção, validação dos exemplos no repositório.
- Interfaces de administração
- Interfaces para o público em geral.

7.7.5 SIEMÊS - Sistema de Identificação de Entidades Mencionadas com Estratégia Siamesa

Participação no HAREM em conjunto com a equipa do pólo do Porto. Implementação do SIEMÊS, um sistema de reconhecimento de entidades mencionadas, usando dois tipos de estratégias. Nesse âmbito, o trabalho consistiu em:

- Análise de métodos estatísticos e implementação destes no SIEMÊS
- Implementação de uma API para aceder e processar os dados do REPENTINO

7.7.6 METRA

No âmbito do Mestrado em Engenharia em Informática, que o bolseiro Luís Cabral frequenta, foi iniciado o desenvolvimento de uma aplicação para análise dos logs do METRA, que é um sistema que invoca vários motores de tradução automática implementado no pólo do Porto e que tem uma grande procura por parte do público, sendo o estudo do comportamento deste interessante para avaliar a qualidade actual da tradução automática de e para o português.

7.7.7 Produção escrita

Elaboração de um relatório a respeito do SAGI, disponível em http://poloclup.linguateca.pt/~lcabral/relatorios/SAGI200409.doc.

Alguma documentação técnica da API do SAGI.

Artigo/documento sobre (o estado da arte n)a identificação de colocações (em progresso).

Apresentação sobre o SAGI no simpósio doutoral a 6 de Maio em Lisboa, disponível em http://poloclup.linguateca.pt/~lcabral/simposio_LC.tgz.

Em colaboração:

Diana Santos, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela & Susana Afonso. "Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa", in Guillermo De Ita Luna, Olac Fuentes Chávez, Mauricio Osorio Galindo (eds.), *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", IX Iberoamerican Conference on Artificial Intelligence (IBERAMIA), November 2004, Puebla, Mexico*, pp. 147-154

Luís Sarmento, Ana Sofia Pinto, Luís Cabral & Débora Oliveira. "REPENTINO - A collaborative gazetteer for Named Entity Recognition in Portuguese", enviado para apreciação.

7.8 Débora Oliveira, bolseira de investigação

O presente relatório visa descrever o trabalho desenvolvido na Linguateca, no âmbito da Bolsa de Investigação em Linguística/Terminologia (Bolsa Busca), de 3 de Julho a 14 de Maio de 2005. O objectivo desta bolsa consiste na melhoria do sistema de pesquisa por palavras (Busca) do sítio da Linguateca, tornando-o mais utilizável e capaz de responder às necessidades e expectativas dos seus utilizadores

7.8.1 Melhoria do Busca

O trabalho desenvolvido no âmbito da melhoria do Busca foi dividido em quatro fases distintas.

1. Utilização do Corpógrafo

As actividades desenvolvidas com esta plataforma incluíram tarefas como a:

- Inserção e pré-processamento dos textos;
- Criação de corpora e de bases terminológicas;
- Extracção de terminologia;
- Pesquisa de definições e de relações semânticas.

Os termos extraídos com o auxílio da ferramenta foram incluídos em bases terminológicas e serviram de base à criação de listas de termos de indexação.

2. Compilação de ficheiros de indexação

Foram compilados três ficheiros de indexação:

1. Termos de indexação seleccionados com base na terminologia extraída;
2. Sinónimos dos termos de indexação compilados no primeiro ficheiro;
3. Possíveis sugestões de pesquisa para cada termo de indexação.

Cada um destes ficheiros foi compilado e progressivamente incrementado e implementado no sistema actual do Busca, em colaboração com o pólo de Oslo.

3. Indexação e Terminologia

Neste contexto, as tarefas desenvolvidas na fase anterior foram temporariamente interrompidas, no sentido de se aprofundar as possíveis relações entre a terminologia e a indexação do sítio. Este trabalho procurou demonstrar como se poderá fazer a selecção de termos de indexação com base em terminologia.

No nosso trabalho fez-se uma análise linguística (morfológica, sintáctica e semântica) e uma análise estatística (análise da frequência, da distribuição e das co-ocorrências) da terminologia extraída com objectivo de ambas contribuirem para uma compilação objectiva de termos de indexação, mas também para uma maior automatização da indexação da informação disponível no sítio.

4. Compilação de uma nova lista de indexação com base em terminologia

Compilou-se uma nova lista de termos de indexação que teve subjacente as conclusões retiradas da análise anterior. Nesta nova lista, os termos de indexação foram divididos por tipo de termo e tipo de resultado esperado pelo utilizador. Esta lista ainda não foi incorporada no sistema de pesquisa, visto que está a ser incrementada.

7.8.2 Auxílio à melhoria do Corpógrafo

- Compilação de padrões de extracção de terminologia (para Alemão);
- Extracção de definições (para Alemão) e de relações semânticas (Português e Inglês);
- Compilação de um corpus de teste sobre Fibromialgia em Português, Inglês e Alemão;
- Sugestão de novas funcionalidades essenciais para a execução do trabalho terminológico.

7.8.3 Auxílio à participação do Pólo no HAREM

- Pesquisa de entidades mencionadas (EMs);
- Criação de categorias de EMs;
- Inserção das EMs no repositório REPENTINO.

7.8.4 Formação

A ida a Oslo (Novembro de 2004) teve como objectivo a participação numa acção de formação sobre os vários projectos da Linguateca, como também a apresentação do trabalho até então realizado relativamente à melhoria do Busca, assim como relativamente à melhoria do sistema de avaliação de Tradução Automática, TrAva. Adicionalmente, durante esta visita, trocaram-se ideias e sugestões sobre o trabalho a desenvolver a partir daí no que se refere ao objectivo principal da Bolsa.

7.8.5 Produção escrita

Débora Oliveira & Ana Sofia Pinto. Extracção de definições no Corpógrafo. Outubro de 2004. Disponível em http://poloclup.linguateca.pt/~doliveira/publico/corpografo.doc.

Débora Oliveira. Relatório sobre o sistema Busca, Outubro 2004. Disponível em http://poloclup.linguateca.pt/~doliveira/publico/Busca.doc.

Débora Oliveira. Relatório sobre o sistema Busca, Dezembro 2004. Disponível em http://poloclup.linguateca.pt/~doliveira/publico/Busca_1.doc.

Diana Santos, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela & Susana Afonso. "Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa", in Guillermo De Ita Luna, Olac Fuentes Chávez, Mauricio Osorio Galindo (eds.), *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", IX Iberoamerican Conference on Artificial Intelligence (IBERAMIA), November 2004, Puebla, Mexico*, pp. 147-154.

Luís Sarmento, Ana Sofia Pinto, Luís Cabral & Débora Oliveira. "REPENTINO - A collaborative gazetteer for Named Entity Recognition in Portuguese", enviado para apreciação.

Débora Oliveira, Luís Sarmento, Belinda Maia & Diana Santos. "Corpus analysis for indexing: when corpus-based terminology makes a difference". In *Corpus Linguistics 2005* (Birmingham, 14-17 July 2005).

7.9 Rosário Morais da Silva, a tempo parcial

As minhas funções no projecto COMPARA incluem:

- o tratamento completo dos textos que farão parte do *corpus* (limpeza ocr, colocação de etiquetas nas várias fases de tratamento dos textos, alinhamento por parágrafo, alinhamento por frase, criação de ficheiros erros e div);
- a realização dos contactos necessários para conseguirmos autorização de utilização dos vários textos;
- a tradução e revisão de novas páginas em português que forem incorporadas no sítio do *corpus*.

No período a que se refere o presente relatório, o meu trabalho pode ser resumido da seguinte forma:

Autorizações pedidas em Portugal (por obra): 6
Autorizações obtidas em Portugal (por obra): 6

Autorizações pedidas a outros países (por obra): 4
Autorizações obtidas outros países (por obra): 1

Textos processados: 11
Textos revistos: 4
Textos traduzidos: 1