

Cooperação Científica e Técnica Luso-Brasileira
ICCTI/CAPES

Anexo Técnico

Projecto DIRPI

Desenvolvimento e Integração de Recursos para Pesquisa de Informação

Participantes:

Portugal:

- Departamento de Informática da Universidade de Évora (DI-UE)
- Departamento de Informática e Centro de Inteligência Artificial da Universidade Nova de Lisboa (UNL)

Brasil:

- Centro de Ciências Exatas e Tecnológicas da Universidade do Vale do Rio dos Sinos (UNISINOS)
- Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

Julho de 2001

1. Instituição Portuguesa

A coordenação do projecto em Portugal será efectuada no Departamento de Informática da Universidade de Évora (DI-UE). Além desta instituição, também colabora neste projecto o Departamento de Informática e o Centro de Inteligência Artificial da Universidade Nova de Lisboa (UNL).

1.1. Antecedentes históricos e contexto institucional

A cooperação científica entre Portugal e Brasil na área da Inteligência Artificial e do Processamento de Língua Natural teve origem, muito provavelmente no primeiro Simpósio Brasileiro de Inteligência Artificial que decorreu em Porto Alegre em 1983, com a participação de diversos investigadores portugueses. Desde então, esta cooperação permitiu, a nível do Rio Grande do Sul, o aparecimento de diversos projectos conjuntos aprovados no âmbito de convénios JNICT/ICCTI/Capes/CNPq. Alguns exemplos, são o projecto MAIS sobre Multi-Agentes e o projecto FUNDAÇÕES sobre Língua Natural, em que estiveram envolvidos nos últimos anos alguns dos proponentes desta proposta (nomeadamente os coordenadores em Portugal e no Brasil). Como consequência desta colaboração foram produzidas diversas publicações em conferências internacionais (ver lista anexa) e várias dissertações de mestrado e doutoramento (mestrado: Sérgio Freitas, Aline Villavicencio; doutoramento: Michael Mora, Flávio Oliveira, Berilhes Garcia).

Por outro lado, e na sequência da investigação efectuada, a colaboração tem vindo a ser alargada a outras instituições, nomeadamente a Universidade de Évora (Portugal) e a UNISINOS (Brasil) – conforme relatórios do projecto FUNDAÇÕES.

Do ponto de vista português, a Universidade de Évora (coordenadora desta proposta em Portugal) tem uma política de cooperação internacional bastante forte, como se pode consultar em <http://www.uevora.pt>. Em concreto, existem de momento 15 projectos de cooperação em diversas áreas entre a Universidade de Évora e universidades brasileiras.

Além deste ponto, o Departamento de Informática da UE organiza desde 1997 em colaboração com a Universidade Nova de Lisboa o Mestrado em Inteligência Artificial Aplicada (MIAA), tendo já mestrando a desenvolver trabalho nesta área de investigação. Está, também, em fase final de aprovação na UE, um novo mestrado mais generalista em Engenharia Informática, que poderá integrar alunos no âmbito desta proposta. Finalmente, existe um programa de doutoramento em Informática da UE que se pretende venha a receber dois doutorandos brasileiros no contexto desta proposta.

1.2. Planos e compromissos institucionais

Em Portugal, este projecto surge na sequência do trabalho que tem vindo a ser realizado nos últimos anos, no âmbito da actividade normal do Departamento de Informática da Universidade de Évora (DI-UE) e do Grupo de Língua Natural (GLINT) da FCT/UNL e do seu Centro de Inteligência Artificial (CENTRIA) (ver projectos referenciados).

Conforme referido anteriormente, a nível académico, o DI-UE e o GLINT têm participado activamente no MIAA-Mestrado em Inteligência Artificial Aplicada organizado em conjunto pela UE e pela UNL. Além desta organização conjunta, o DI-UE irá iniciar no ano lectivo de 2001/2002 o Mestrado em Engenharia Informática da UE, enquanto o GLINT já tem participado activamente no Mestrado

em Engenharia Informática da Universidade Nova de Lisboa. É de salientar que as duas disciplinas de Processamento de Língua Natural do MIAA são asseguradas pelos dois elementos da UE – Paulo Quaresma e Irene Rodrigues – e que Gabriel Lopes é responsável por uma disciplina do Mestrado em Engenharia Informática da UNL sobre extracção de informação em textos.

Em termos de projectos de cooperação, o DI-UE já participou activamente no projecto ICCTI/CAPES FUNDAÇÕES e tem em fase de avaliação o projecto COOPERATIVE, submetido à FCT, que visa desenvolver trabalho na área deste projecto e que poderá vir a servir como enquadramento do trabalho dos doutorandos brasileiros.

Finalmente, existe o apoio institucional explícito por parte do DI-UE, que pretende vir a estender e consolidar a sua actividade, e da própria UE, ao nível do seu Vice-Reitor para a cooperação internacional.

1.3.Motivação técnico-científica

A investigação na área de PLN, e especificamente na Língua Portuguesa, é um tema de interesse estratégico para os Países de Língua Oficial Portuguesa. Neste contexto, o desenvolvimento de ferramentas especializadas para o Português e o estabelecimento de programas de formação avançada nesta área representam um factor importante para a consolidação e incremento do papel da Língua Portuguesa no contexto de um mundo cada vez mais globalizado.

No conjunto vasto de problemas a enfrentar em PLN, assume particular importância a necessidade de lidar automaticamente com conjuntos vastos de textos, tendo a capacidade de pesquisar e extrair a informação que eles contêm. A linha de investigação em que se enquadra esta proposta pretende apoiar o desenvolvimento de ferramentas nas seguintes áreas (descrição pormenorizada na secção 3):

- Etiquetadores morfossintácticos;
- Analisadores sintácticos;
- Analisadores de co-referências
- Anotadores de corpora
- Gestão de diálogos;
- Integração das ferramentas desenvolvidas num sistema inteligente de pesquisa de informação

Com a presente proposta, será possível conjugar esforços entre os diversos grupos de investigação, obter resultados mais rapidamente e avaliar as soluções propostas para o Português de Portugal e do Brasil.

Este factor, permitirá criar as condições necessárias para perspectivar a participação conjunta em projectos de investigação.

2. Unidade Executora (Portugal)

A coordenação do projecto em Portugal estará a cargo do Departamento de Informática da Universidade de Évora (DI-UE), existindo ainda a participação da Universidade Nova de Lisboa, através do seu Departamento de Informática e do Centro de Inteligência Artificial.

2.1. Breve história do DI-UE e do DI-UNL

A Universidade de Évora foi criada em 1559 durante a regência de D. Catarina da Áustria e confirmada por Bula Papal de Paulo IV. No século XVIII, no âmbito das reformas políticas e culturais realizadas pelo Marquês de Pombal, a Universidade foi extinguida, tendo sido re-criada somente em 1973. A área científica de Informática enquadrou-se inicialmente no departamento de Matemática, tendo somente em 1999 sido criado o departamento de Informática. A licenciatura em Engenharia Informática foi lançada em 1995, estando de momento a decorrer o processo de acreditação na Ordem dos Engenheiros. Para além disso, o DI-UE co-organiza o Mestrado em Inteligência Artificial Aplicada desde 1997 e vai iniciar o seu próprio Mestrado em Engenharia Informática a partir do corrente ano.

O Departamento de Informática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa foi criado em 1974, tendo como objectivo realizar o ensino na área de Informática. A licenciatura em Engenharia Informática iniciou-se em 1975, sendo a primeira do género em Portugal. Desde então, o DI-UNL tem vindo a consolidar e a aumentar as suas actividades, como se pode observar a partir da avaliação dos seus centros de investigação na área de informática: CENTRIA e CITI.

2.2. Organização e Direcção do DI-UE e DI-UNL

O DI-UE está inserido na Área Departamental de Ciências Exactas, em conjunto com os departamentos de Matemática, Física e Química. Como órgãos constituintes, existe o Conselho do Departamento, constituído pelos professores e por representantes dos docentes não doutorados. O Presidente do Conselho do Departamento, actualmente a Prof. Irene Rodrigues, tem como competências preparar as reuniões do Conselho e executar as suas deliberações; assegurar a gestão dos meios humanos e materiais; preparar convénios, acordos e contratos de prestações de serviços; e zelar pela conservação e manutenção das instalações e bens do departamento.

O DI-UNL constitui uma unidade orgânica de ensino graduado e pós-graduado e de investigação na área da Informática. O departamento possui autonomia pedagógica, científica e de gestão. Como órgãos constituintes, existe o Conselho do Departamento, constituído por membros permanentes (professores, investigadores e técnicos superiores doutorados) e por membros não permanentes (representantes dos docentes, investigadores e técnicos superiores não doutorados). O Presidente do Conselho do Departamento é actualmente o Prof. Pedro Barahona. O DI-UNL está, ainda, dividido em três secções: Ciência e Tecnologia da Programação; Arquitectura de Sistemas de Computadores; Sistemas Simbólicos de Decisão e de Informação.

2.3. Quadro de Investigadores envolvidos com o projecto

Paulo Quaresma (UE) – Professor Auxiliar do DI–UE. Director da Comissão de Curso da Licenciatura em Engenharia Informática. Presidente da Comissão Científica do DI–UE. Coordenará a equipa portuguesa, discutindo e integrando as diversas abordagens propostas, tendo uma participação mais activa na área da gestão de diálogos.

Irene Pimenta Rodrigues (UE) – Professor Auxiliar do DI–UE. Presidente do Conselho do Departamento. Coordenará a área de pesquisa de informação e de co-referências.

Gabriel Pereira Lopes (UNL) – Investigador Principal do DI–UNL. Como consequência da experiência acumulada na área e em projectos de cooperação similares, poiará na coordenação e integração do trabalho desenvolvido.

Nuno Cavalheiro Marques (UNL) – Investigador do CENTRIA. De momento, Professor Auxiliar da Universidade Aberta, sendo provavelmente a partir de Outubro, docente da UNL. Coordenará a área de etiquetagem morfossintáctica.

Vitor Rocio (UNL) – Investigador do CENTRIA. Em princípio, defenderá a tese de doutoramento até final de 2001. Coordenará a área de análise sintáctica.

Alexandre Agustini (UNL) – Doutorando da UNL. Trabalhará na área de anotação de corpora.

Pablo Gamallo (UNL) – Bolseiro Pós–PhD da UNL. Coordenará a área de anotação de corpora.

Joaquim Ferreira da Silva (UNL) – Assistente do DI–UNL. Trabalhará na área de pesquisa de informação.

Antônio Ribeiro (UNL) – Investigador do CENTRIA. Em princípio, defenderá a tese de doutoramento até ao primeiro semestre de 2002. Trabalhará na área de pesquisa de informação

2.4. Infra-estrutura

O DI–UE possui as infra-estruturas necessárias ao desenvolvimento do trabalho proposto para os seus investigadores (docentes e alunos de mestrado e/ou doutoramento). Nomeadamente, a nível de computadores, redes informáticas, impressoras e gabinetes. Existe, ainda, apoio de secretariado e uma biblioteca da área de Ciências Exactas.

O CENTRIA da UNL disponibiliza para os seus investigadores e alunos de pós-graduação as infra-estruturas necessárias ao seu desempenho (computadores, gabinetes, redes, etc.). A Faculdade de Ciências e Tecnologia da UNL possui ainda uma biblioteca bastante bem equipada na área de Inteligência Artificial e de PLN.

2.5. Indicadores de Desempenho (Produção Científica)

Lista de artigos mais significativos para o projecto:

- [**BalsaLopes2000**] Balsa, J. and Lopes, J.G.P. 2000. "A Distributed Approach for a Robust and Evolving NLP System". in Dimitis Christodoulakis (Ed.) "Natural Language Processing – NLP 2000, Second International Conference, Patras, Greece, June 2000, Proceedings". Lecture Notes in Artificial Intelligence Springer-Verlag 1835. pp. 151–161.
- [**Diasetal2000**] Dias, G. & Guilloire, S. & Bassano, J.C. & Lopes, J.G.P.(2000). Extraction Automatique d'unités Lexicales Complexes: Un Enjeu Fondamental pour la Recherche Documentaire. In *Journ. Traitement Automatique des Langues*, Vol 41:2, Christian Jacquemin (ed.). Paris, France.
- [**Gamalloetal2001a**] Gamallo P., Gasperin C. Agustini A. Lopes G.P. (2001a) "Syntactic-Based Methods for Measuring Word Similarity". *TSD-2001*, Plenz, CZ. (to appear).
- [**Gamalloetal2001b**] Gamallo P., Gasperin C. Agustini A. Lopes G.P. Lima V.L (2001b) "The use of syntactic context for measuring word similarity", *ESSLLI-2001*, workshop on Semantic Knowledge Acquisition and Categorisation, Helsinki, CZ. (to appear).
- [**Gamalloetal2001c**] Gamallo P., Agustini A., Lopes G.P: (2001) "The role of co-specification for the acquisition of selection restrictions from unsupervised corpora". *AFIA-2001, workshop A3CTE*, Grenoble, France (27–33).
- [**LopesMarques1994**] José Gabriel Lopes, Nuno Marques. *Reconhecimento de Neologismos*. In Hector Geffner, editor, *Actas del IV Congreso Iberoamericano de Inteligencia Artificial Y VII Conferencia Nacional sobre Inteligencia Artificial (IBERAMIA'94)*, páginas 271–283, Caracas, Venezuela. McGraw-Hill Venezuela. 1994.
- [**Lopesetal1994**] Lopes, J. G.; Marques, N. C.; Rocio, V. J. *POLARIS: Portuguese Lexicon Acquisition and Retrieval Interactive System*, aceite na conferência The Practical Applications of Prolog na Royal Society of Arts, página 665, Londres-Reino Unido, Abril 26–29. 1994.
- [**Lopes1998**] Lopes, J. (1998). Desenvolvimento de dicionários electrónicos e de bases de dados lexicais para processamento computacional do português, Publicado na revista sobre estudos hispânicos da Universidade de Augsburg.
- [**LopesRocioBalsa1999**] J. Gabriel Pereira Lopes, Vitor Rocio and João Balsa da Silva "Superando a incompletude da informação lexical".in: P. Marrafa e M. A. Mota(eds.). *Linguística Computacional: Investigação Fundamental e Aplicações*. Lisboa: Edições Colibri, ISBN 972-772-090-0.
- [**Marques2000**]. Nuno Marques "Uma Metodologia para a Modelação Estatística da Subcategorização Verbal". Dissertação de Doutoramento. Universidade Nova de Lisboa. 2000.
- [**MarquesLopes2001**] Nuno Marques, José Gabriel Lopes e Carlos Agra Coelho. *A POS-Tagger Generator for Unknown Languages*. Actas SEPLN 2001. Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural. 2001.
- [**MarquesLopes2001b**] Nuno Marques e José Gabriel Lopes. *Tagging With Small Training Corpora*. In *Proceedings of the International Conference on Intelligent Data Analysis*, Lecture Notes in Artificial Intelligence Series, Cascais, Portugal. Springer Verlag. Setembro 2001.
- [**PoesioVieira1998**] Poesio, M., Vieira, R. (1998). Corpus-based investigation of definite description use. *Computational Linguistics*, Vol. 24 (2): 183–216.
- [**Quaresma1997**] Paulo Quaresma. Inferência de Atitudes em Diálogos. Dissertação de Doutoramento. Universidade Nova de Lisboa. 1997.

- [QuaresmaRodrigues2001] Quaresma, P. & Rodrigues, I. (2001). "Using Logic Programming to model Multi-Agent Web Legal Systems – An Application Report". In ICAIL'2001 International Conference on AI & LAW, St. Louis, EUA, to be published by ACM, 2001.
- [QuaresmaRodrigues2000] Paulo Quaresma and Irene Pimenta Rodrigues. 2000. "Automatic classification and Intelligent clustering for WWW information retrieval systems", In: Journal of Information Law and Technology (JILT), vol. 2, <http://elj.warwick.ac.uk/jilt/00-2/> (Extended and revised version of the BILETA'2000 paper).
- [QuaresmaRodrigues2000b] Paulo Quaresma and Irene Pimenta Rodrigues. 2000. "Construire le contexte d'une interface Web pour des bases de données juridiques", In: D. Bourcier and P. Hassett and C. Roquilly (ed.). "Droit et Intelligence Artificielle. Editions Romillat: Paris, France. ". p. 46–58. Collection Droit et Technologies, ISBN 2–87894–017–2.
- [QuaresmaRodrigues2000c] Paulo Quaresma and Irene Pimenta Rodrigues. 2000. "A Dialogue Manager for a WWW-Based Information Retrieval System", In: H. Larsen and J. Kacprzyk and S. Zadrozny and T. Andreassen and H. Christiansen (eds.). "FQAS 2000: Fourth International Conference on Flexible Query Answering, Warsaw, Poland, October, 2000. Proceedings. Lecture Notes on Advances in Soft Computing. Physica-Verlag, Heidelberg, New York (ISBN 3–7908–1347–8). pp. 592–601.
- [QuaresmaRodrigues2000d] Paulo Quaresma and Irene Pimenta Rodrigues. 2000. "Cooperative Information Retrieval Dialogues Through Clustering", In: Petr Sojka and Ivan Kopeck and Karel Pala (eds.). TSD 2000 – Third International Workshop on TEXT, SPEECH and DIALOGUE, Brno, Czech Republic, September, 2000, Proceedings. Lecture Notes in Artificial Intelligence LNCS/LNAI 1902. Springer-Verlag. pp. 415–420. Proceedings preliminary WWW page: <http://www.fi.muni.cz/tsd2000/tsdproc.html>.
- [QuaresmaRodrigues2000e] Paulo Quaresma and Irene Pimenta Rodrigues. 2000. "A Logic Programming Agent for Controlling Distributed Web Dialogues". In: Osamu Yoshie (ed.). "INAP'2000 – The 13th International Conference on Applications of Prolog, Tokyo, Japan, October 2000". Tokyo: Ren Associates Inc. (ISSN 1345–0980). pp. 46–53.
- [Ribeiroetal2000a] António Ribeiro, Gabriel Lopes e João Mexia, "Linear Regression Based Alignment of Parallel Texts Using Homograph Words". In Werner Horn (ed.) (2000) ECAI 2000: Proceedings of the 14th European Conference on Artificial Intelligence, volume 54, © IOS Press 2000, Amsterdam, The Netherlands, pp. 446–450. Berlin, Alemanha, 2000 Agosto 20–25.
- [Ribeiroetal2000b] António Ribeiro, Gabriel Lopes e João Mexia, "Using Confidence Bands for Parallel Texts Alignment". In Proceedings of the 38th Conference of the Association for Computational Linguistics (ACL 2000), © Association for Computational Linguistics 2000, pp. 432–439. Hong Kong, China, 2000 Outubro 3–6.
- [Ribeiroetal2000c] António Ribeiro, Gabriel Lopes e João Mexia (2000c) "A Self-Learning Method of Parallel Texts Alignment". In John White (ed.) (2000), Envisioning Machine Translation in the Information Future – Proceedings of the 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000 – Lecture Notes in Artificial Intelligence, volume 1934, © Springer-Verlag, Berlin, Alemanha, pp. 30–39. Cuernavaca, México, 2000 Outubro 10–14.
- [Rocio2001] Vitor Rocio. "Infra-estrutura para superação de falhas no processamento sintático de textos", Dissertação de Doutoramento. Universidade Nova de Lisboa. Entregue para avaliação em Julho 2001.
- [Rocio2001b] V. Rocio, E. de la Clergerie, J.G.P.Lopes (2001). "Tabulation for multi-purpose partial parsing". Journal Grammars. Volume 4 Number 1. Kluwer. pag. 41–65.

- [Silvaetal1999] J.F. Silva, Gael Dias, Sylvie Guilloiré, José Gabriel P. Lopes. 1999. "Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units". In: P. Barahona (ed.) *Progress in Artificial Intelligence: 9th Portuguese Conference on AI, EPIA'93, Évora Portugal, September 1999, Proceedings*. Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol. 1695, p. 113-132 (1999).
- [SilvaLopes1999] J.F. da Silva, and J.G.P.Lopes. 1999. "A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora". In *Proceedings of the Sixth Meeting on Mathematics of Language (MOL6)*, Orlando, Florida July 23-25, 1999.}. pp. 369-381
- [SilvaLopes1999b] J.F. da Silva, and J.G.P.Lopes. 1999. "Extracting Multiword Terms from Document Collections". In *Proceedings of the VExTAL: Venezia per il Trattamento Automatico delle Lingue*, November 22-24, 1999.}.
- [Silvaetal1999b] J.F. da Silva, J.G.P.Lopes, M. F. Xavier, and G. Vicente. 1999. "Relevant Expressions in Large Corpora". In Anne Condaminés, Cécile Fabre et Marie-Paule Péry-Woodley (eds.) *Actes de l'atelier "Corpus et Traitement Automatique des Langues: Pour une réflexion méthodologique" (TALN'99)*, Institut d'Etudes Scientifiques, Cargèse, Corse (France), July 12-17.}. Pp. 86-94. Published by ATALA.
- [Vilavencioetal95] A. Vilavencio, N. Marques, G. Lopes, F. Vilavencio. *Part-of-Speech Tagging for portuguese Texts*. In Jacques Wainer e Ariadne Carvalho, editores, *Advances in Artificial Intelligence: Proceedings of the XII Brazilian Symposium on Artificial Intelligence*, Lecture Notes in Artificial Intelligence 991, paginas 323-332, Campinas, Outubro 11-13. Springer Verlag. 1995.

3. Plano de Trabalho

3.1. Objectivos

Com a realização deste projeto de cooperação, tem-se por objetivo compartilhar os resultados da pesquisa na área realizada por grupos de processamento de linguagem natural no Brasil, em especial na região sul, e por grupos portugueses. Este projecto permitirá consolidar e aprofundar as ligações internacionais do DI-UE e do GLINT com a UNISINOS e PUCRS, simultaneamente através do desenvolvimento de novos produtos e exploração das ferramentas já construídas, bem como consolidar, reforçar e estender a novas áreas e investigadores as ligações já existentes e que foram apoiadas no âmbito dos projectos NALAMAS e ICCTI/CAPES FUNDAÇÕES (a terminar em 2001).

Em concreto, pretende-se:

- Efectuar a integração dos diversos investigadores dos dois países em tarefas concretas de projectos comuns existentes nas instituições proponentes. Estas tarefas, bem como os investigadores que lhes irão ser associados, estão descritas na fundamentação teórica desta proposta. Gostaríamos de salientar que esta integração irá permitir um desenvolvimento natural das actividades em andamento entre as diversas entidades intervenientes.
- Continuar e estender o fluxo de de formação entre os dois países nestas áreas de investigação. Este fluxo foi iniciado com o projecto FUNDAÇÕES e pretende-se que seja consolidado com a extensão aos programas de doutoramento da Universidade de Évora e à UNISINOS.

- Reforçar a apresentação de cursos de curta duração e palestras relacionadas com a actividade associada às visitas, reuniões dos pesquisadores com alunos de graduação e pós-graduação envolvidos em actividades directamente ligadas ao projeto ou actividades relacionadas e com pesquisadores da instituição visitada.
- Potenciar o desenvolvimento de projectos de investigação comuns, a ser apresentados em programas futuros.

3.2. Justificações

Pesquisas na área de Processamento de Linguagem Natural (PLN) têm demonstrado seus efeitos na indústria de software, tendo como alguns exemplos o desenvolvimento de corretores gramaticais, os tradutores multilíngues e sistemas de recuperação de informação. Ferramentas e recursos para o tratamento computacional da língua portuguesa são bastante escassos em relação a outras línguas como inglês, espanhol, francês e japonês, apesar do número de usuários em potencial ser significativo.

Há entre grupos de pesquisa brasileiros e portugueses o interesse comum na produção de trabalhos que contemplem o desenvolvimento de tecnologias para o processamento da língua Portuguesa.

Com a realização deste projeto de cooperação, tem-se por objetivo consolidar, integrar, e adaptar recursos obtidos através da pesquisa realizada na área por pesquisadores brasileiros, em especial da região sul, e pesquisadores portugueses.

No Estado do Rio Grande do Sul, esta linha de pesquisa possui grupos já formados (PUCRS, mestrado em ciência da computação com linha de pesquisa em PLN) e grupos em formação (UNISINOS, mestrado em Lingüística Aplicada criado em 2000 e mestrado em Computação Aplicada criado em 1999).

Em Portugal, este projecto surge na sequência do trabalho que tem vindo a ser realizado nos últimos anos, no âmbito da actividade normal do Departamento de Informática da Universidade de Évora (DI-UE) e do Grupo de Língua Natural (GLINT) da FCT/UNL e do seu Centro de Inteligência Artificial (CENTRIA) (ver projectos referenciados). A nível académico, o DI-UE e o GLINT têm participado activamente no MIAA-Mestrado em Inteligência Artificial Aplicada organizado em conjunto pela UE e pela UNL. Além desta organização conjunta, o DI-UE irá iniciar no ano lectivo de 2001/2002 o Mestrado em Engenharia Informática da UE, enquanto o GLINT já tem participado activamente no Mestrado em Engenharia Informática da Universidade Nova de Lisboa. A nível de projectos de cooperação, existe um histórico vasto de projectos conjuntos entre as diversas entidades participantes nesta proposta

Este projecto permitirá consolidar e aprofundar as ligações internacionais do DI-UE e do GLINT com a UNISINOS e PUCRS, simultaneamente através do desenvolvimento de novos produtos e exploração das ferramentas já construídas.

3.3. Metas

Os vários projetos em desenvolvimento pelos grupos envolvidos na cooperação irão se beneficiar da troca de experiência, conhecimentos e sistemas desenvolvidos por variados grupos. As visitas de intercâmbio possibilitarão:

- a realização de cursos em áreas específicas relacionadas aos projetos, tais como, resolução de anáfora pronominal e resolução de descrições definidas na língua portuguesa, métodos estatísticos em processamento de linguagem natural, recuperação de informação e gestão de diálogos para recuperação de informação;
- apresentação dos resultados obtido pelas investigações em desenvolvimento:

- compartilhamento e formação em ferramentas e recursos, que possibilitem a construção de corpora anotados (morfologicamente, sintacticamente e semanticamente) do português.

Em termos mais globais, tem-se como meta final a ampliação da colaboração existente actualmente entre as diversas entidades participantes nesta proposta, permitindo o reforço do número de investigadores dos dois países envolvidos nestas áreas de investigação.

Em termos específicos, tem-se como meta orientar 2 alunos de mestrado da UNISINOS e/ou PUC na área e outros 2 alunos de mestrado na UE e UNL; encaminhar um doutorado sandwich (PUC) e preparar um aluno brasileiro para que ao terceiro ano do projeto tenha condições de integrar-se aos grupos de pesquisas portuguesas para formação ao nível de doutoramento.

3.4. Metodologia

Pretende-se efectuar a integração dos diversos investigadores dos dois países em tarefas concretas de projectos comuns existentes nas instituições proponentes. Estas tarefas, bem como os investigadores que lhes irão ser associados, já foram descritas na fundamentação teórica desta proposta. É de salientar que esta integração irá permitir um desenvolvimento natural das actividades em andamento entre as diversas entidades intervenientes.

As visitas constarão de apresentação de cursos de curta duração e palestras relacionadas com a atividade associada às visitas, reuniões dos pesquisadores com alunos de graduação e pós-graduação envolvidos em atividades diretamente ligada ao projeto ou atividades relacionadas e com pesquisadores da instituição visitada.

Da parte brasileira, pretende-se o tanto quanto possível a integração de alunos de graduação, e pós-graduação em orientação pelos professores pesquisadores participantes com as atividades do projeto.

3.5. Conteúdo básico – Fundamentação teórica

A presente proposta baseia-se nos projetos e pesquisas em desenvolvimento pelos grupos brasileiros e portugueses. A fundamentação teórica, a organização dos vários projetos e o relacionamento entre suas partes são descritas a seguir. Primeiramente são listadas as condições atuais das partes e em seguida a perspectiva futura de trabalho para dada uma.

3.5.1. Projetos e pesquisas atuais

– Construção de etiquetadores para textos da língua portuguesa

O trabalho com etiquetadores morfossintáticos, visando a construção de desambiguadores de etiquetas de carácter sintático atribuídas a uma palavra, iniciou-se com o projecto DIXIT (ver lista de projectos abaixo para mais detalhes sobre os diversos projectos) descrito em [Vilavicencioetal95], tendo sido continuado no âmbito do trabalho com o Centro de Estudos Comparados, da Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa (projectos CESA e ESA) e PGR, tendo sido um dos trabalhos base da tese [Marques2000]. Os resultados mais recentes sobre este sistema podem ser encontrados nos artigos [MarquesLopes2001] e [MarquesLopes2001b].

– **Construção de analisadores sintáticos**

Este trabalho resulta da pesquisa efectuada no âmbito de [Rocio2001]. O objectivo é a análise sintáctica de texto não restrito. A análise total do texto é, em geral, inviável devido a vários fatores, entre outros, destaque-se a ocorrência de erros no texto de entrada ou etiquetado e a incompletude dos dicionários e gramáticas disponíveis. No entanto, uma análise apenas parcial da entrada é possível. Um sistema de diagnóstico e correcção de falhas permite criar agentes de recuperação que têm por objectivo tentar melhorar os resultados da análise [LopesRocioBalsa99], [BalsaLopes2000]. Neste trabalho utilizam-se técnicas de tabulação do sistema DyALog para possibilitar uma análise eficiente do texto [Rocio2001b].

– **Construção de sistemas para resolução de correferência**

O projeto ANACORT (UNISINOS, PUCRS), em fase de conclusão, visou o tratamento da correferência de descrições definidas em textos da língua portuguesa. Foi preparado um corpus para a pesquisa que contém a identificação das ocorrências de sintagmas nominais em textos da língua portuguesa do Brasil, mais especificamente artigos do Jornal Correio do Povo de Porto Alegre. O corpus foi analisado manualmente e os algoritmos de resolução de correferência estão em fase de testes [Vieiraetal2000]. Este trabalho baseia-se em resultados obtidos por estudos realizados sobre o mesmo problema em relação à língua inglesa ([PoesioVieira1998]; [Vieira1998], [VieiraPoesio2000a]; [VieiraPoesio2000b], [VieiraPoesio2000c]).

O projeto MLXML (UNISINOS), em fase inicial, visa a construção de esquemas padronizados para a anotação de informação lingüística e semântica em corpus (textos e documentos web), de acordo com padrões que têm sido propostos recentemente por entidades representativas da área (EAGLES, ELSNET, XCES, ???).

O projeto COMMON-REFs (UNISINOS, PUCRS/ ProTeM-CC-Inria), em fase inicial, visa o estudo de expressões referenciais, entre elas a correferência em textos e diálogos multilíngues.

– **Sistemas de pesquisa de informação em textos da língua portuguesa:**

O trabalho realizado nesta área é, essencialmente, resultado do projecto PGR (ver lista de projectos) que permitiu a construção de um sistema de pesquisa de informação para textos em Língua Portuguesa. Entre outras características, este sistema incorporou num motor de pesquisa para bases de textos, uma base lexical de 900,000 palavras (POLARIS) que permite efectuar pesquisas tendo em conta as diversas formas das palavras. O sistema possui, ainda, a possibilidade de extrair automaticamente e utilizar nas pesquisas um conjunto de expressões relevantes existentes nos textos. Este processo é efectuado recorrendo a métodos estatísticos e medidas de informação mútua que são independentes da Língua (Ferreira da Silva et al 1999a, 1999b).

Por outro lado, o projeto MLXML (UNISINOS) investiga o projeto semântico de páginas web e a busca de informação em tais contextos.

– **Gestão de diálogos para interação com sistemas de pesquisa de informação**

Esta área foi desenvolvida inicialmente no âmbito da tese [Quaresma1997], em que foi proposto um ambiente de programação em lógica para a modelação de agentes computacionais com a capacidade de participar de um modo activo, cooperativo e inteligente em diálogos. Este trabalho tem vindo a ser extendido em diversos artigos publicados em conferências internacionais (Quaresma e Rodrigues, 2001, 2000) e foi parcialmente implementado no sistema desenvolvido no âmbito do projecto PGR.

– Corpora e construção de corpora anotados

Anotação morfológica

A anotação morfológica do corpus foi até agora realizada sobre 3 tipos distintos de corpora (projecto CORPUS): o corpus de notícias da agência LUSA, onde foram anotados automaticamente mais de 40 milhões de palavras com base em apenas 5 000 palavras corrigidas manualmente. O corpus PGR, onde foi construído um corpus com 20.000 palavras corrigidas manualmente e, no âmbito dos projecto ESA e CESA, sobre um corpus de português antigo (séculos XIII a XIX), onde já foram etiquetadas e corrigidas manualmente mais de 50.000 palavras.

Anotação sintáctica parcial

A anotação sintáctica parcial tem sido efectuada tendo como base a anotação morfológica. Até agora, temos disponíveis textos do corpus PGR, corpus agência notícias LUSA e FCSH analisados parcialmente de forma automática. Estes textos contêm ainda erros de análise, pois não foram corrigidos manualmente.

Anotação semântica

Através do projeto MLXML, estão sendo elaborados modelos e esquemas de anotação lingüística de corpus. O projeto COOMOn-REFs focaliza no estudo do comportamento das expressões referenciais e as consequências deste na elaboração de esquemas de anotação de corpus.

Corpus paralelo

Um outro trabalho realizado com corpus é o alinhamento de textos paralelos com objectivo de extrair equivalentes de tradução. Este trabalho teve como objectivo a construção de dicionários bilingues de traduções de palavras ou expressões (multiword units) a partir de textos que são traduções mútuas (textos paralelos). Foram desenvolvidas metodologias para identificar as traduções de partes de textos paralelos – alinhamento – (Ribeiro et al., 2000a,b) e para identificar as expressões equivalentes [Ribeiroetal2000c]. Este trabalho foi explorado com textos paralelos do Jornal Oficial das Comunidades Europeias nas onze línguas oficiais da União Europeia e de textos da Imprensa Oficial de Macau com textos em Português, Inglês e Chinês.

Extracção de informação semântica em corpora

A extracção automática de informação semântica tem sido efectuada em cima de texto analisado parcialmente (e portanto também etiquetado com anotação morfológica) da Procuradoria Geral da República. A metodologia adoptada baseia-se em duas estratégias: Por um lado, é usada uma variante da metodologia de Gregory Grefenstette [Grefenstette1994] por forma a construir um thesaurus com palavras semanticamente relacionadas (Gamallo et al., 2001ab). Por outro lado, tem sido elaborado um sistema parcialmente inspirado na abordagem de David Faure e Claire Nédellec ([FaureNédellec1998]) para a extracção de classes de subcategorização semântica ou restrições de selecção [Gamalloetal2001c].

3.5.2.Projetos e pesquisas a serem desenvolvidos em cooperação

– Construção de lematizadores

A separação de um corpus nas suas palavras é um trabalho necessário a qualquer análise de texto. Quando essa análise é efectuada de forma totalmente automática aumentam as possibilidades de erro. Pretende-se investigar as possibilidades de melhorar os sistemas actualmente disponíveis para lematização de corpora por forma a minimizar esta hipótese de erro. De crucial importância para o resto do projecto será o facto de o sistema de lematização automática poder ter em conta os recentes trabalhos sobre extracção de unidades lexicais multi-palavra (Ferreira da Silva et al

1999a, 1999b, [Diasetal2000]) e as ferramentas disponíveis para tratamento da norma XML (<http://www.w3c.org/XML>).

Este trabalho será efectuado tendo por base a experiência sobre lematização automática do GLINT, nomeadamente, a experiência resultante do trabalho com as primeiras versões do etiquetador morfossintáctico, no âmbito dos projectos ESA, CESA e PGR. Tentar-se-á desenvolver este módulo com recurso a projectos de fim de curso na área de informática e com a colaboração de um estudante de Mestrado. Caso o projecto seja financiado, este trabalho poderá decorrer ou na UNISINOS ou na PUC-RS.

– Adaptação dos etiquetadores de textos em português de Portugal para a língua portuguesa do Brasil

O etiquetador descrito em (Marques e Lopes 2001a, 2001b) tem como grande vantagem o facto de necessitar de um volume muito reduzido de texto marcado para a aprendizagem automática de desambiguação de textos de tipo distinto (este valor ronda as 10 000 palavras marcadas). O método a utilizar deverá ser semelhante ao descrito para a construção de corpora anotados (ver abaixo). Note-se que as etiquetas a utilizar deverão ter em conta quer as necessidades do analisador sintáctico quer as necessidades de pesquisa de informação.

– Dicionários para a língua portuguesa do Brasil

A base de dados lexical (POLARIS) existente actualmente foi construída com base em dicionários para a Língua Portuguesa na sua *versão* de Portugal (Lopes et al. 1994, [Lopes1998]). Existe a necessidade de adaptar esta base de dados à Língua Portuguesa do Brasil. Para tal, dever-se-á alterar a definição da base de dados de modo a permitir a co-existência de várias *versões*. Posteriormente, pretende-se vir a fazer a aquisição automática de vocabulário a partir de textos em Português do Brasil e completar o dicionário tentando estabelecer protocolos com entidades brasileiras que possuam tal informação.

– Construção de analisadores sintácticos

O objectivo é realizar a migração do ambiente de análise sintáctica e sistema de diagnóstico desenvolvidos por Vitor Rocio ao português do Brasil. Os sistemas foram desenvolvidos em PROLOG e utilizam o sistema DyALog (Linux), de forma que a implementação deve ocorrer de forma directa. Os sistemas devem ser adaptados, contudo, para executar sobre os dicionários disponíveis para a nova língua, e os vários níveis de especificação gramatical utilizados devem ser avaliados e adaptados para o português brasileiro.

– Utilização da resolução de correferência em pesquisa de informação

Investigar as relações entre expressões referenciais, expressões correferentes e expressões relevantes em textos da língua Portuguesa, com o objetivo de obter melhorias em busca automática de informação em textos escritos.

– Utilização de resolução de correferência em sistemas de gestão de diálogo

Nesta tarefa pretende-se utilizar os resultados da resolução de correferência no contexto de sistemas de gestão de diálogos. Nomeadamente, pretende-se vir a suportar diálogos em que os utilizadores (usuários) possam utilizar frases em que existam correferências entre as diversas frases interrogativas (por exemplo, relações anafóricas inter-sentenças).

– Corpora e construção de corpora anotados

Anotação morfológica

Utilizando o etiquetador morfossintáctico treinado sobre o texto disponível em português de Portugal, será etiquetado automaticamente um volume de texto até 10000 palavras em português do Brasil. Este texto será corrigido manualmente. Serão etiquetadas mais 50 000 palavras de texto *novo*. Sobre este texto serão apenas corrigidas as palavras *desconhecidas* no léxico construído com recurso ao corpus de 10 000 palavras (segundo as estimativas actuais, aproximadamente 5000 palavras). Com este processo será esperada um grau de correcção de 96% sobre texto novo.

Anotação sintática parcial

A disponibilidade de corpora com anotação sintática é fundamental para um grande conjunto de aplicações. A partir da adaptação do analisador sintáctico desenvolvido pelo grupo GLINT para o português do Brasil será seleccionado um pequeno corpus de treino para avaliar tanto a adequação das gramáticas desenvolvidas quanto o tipo de anotação sintática desejada. Após estes "ajustes" o sistema poderá ser utilizado para realizar a anotação do texto previamente etiquetado (morfo-sintaticamente).

Anotação semântica

A partir de um corpus anotado parcialmente quer-se trabalhar na avaliação de possibilidade de anotação semântica de corpus em relação aos processos de referência e co-referência, com base em trabalhos desenvolvidos para o inglês e o português ([Vieira2000]; Vieira e Poesio, 1998, 2000a, 2000b, 2000c)

Corpus paralelo

Um dos problemas identificados na extracção de equivalentes de tradução foi a tradução de expressões com verbos. A flexão dos verbos reduz a possibilidade de serem facilmente e correctamente identificadas as traduções dos verbos noutras línguas. A lematização dos textos poderá melhorar a extracção dos equivalentes de tradução destas expressões. O dicionário bilingue de traduções extraído permite resolver alguns problemas de traduções de textos.

Estudo de correferência em corpus paralelo

Um dos outros problemas identificados e que será interessante explorar é a variação do comportamento das co-referências nas diferentes línguas. Tanto quanto conhecemos, não existe nenhum trabalho de análise comparativa deste problema. Seria interessante explorá-lo para os pares Português-Inglês, Português-Francês e Português-Espanhol para o qual já existe projetos em desenvolvimento (COMMON-REFs).

Extracção de informação semântica de corpora

- 1) O thesaurus extraído automaticamente pode ser utilizado em sistemas de pesquisa de informação para expandir as interrogações, e em sistemas de resolução de correferências. A integração do thesaurus nestes sistemas e a sua influência nos resultados obtidos permitirá avaliar a pertinência das relações thesáuricas extraídas.
- 2) As restrições de selecção extraídas automaticamente serão integradas no dicionário POLARIS. Esta nova informação semântica será usada pelo analisador, permitindo a resolução de um grande número de dependências sintácticas.

3.6. Pessoal envolvido no Brasil e em Portugal (CV em anexo)

Portugal

Docentes Doutorados

Paulo Quaresma (UE)

Irene Pimenta Rodrigues (UE)

Nuno Cavalheiro Marques (UNL)

Investigadores Doutorados

Gabriel Pereira Lopes (UNL)

Pablo Gamallo (UNL)

Doutorandos Mestres

Vitor Rocio (UNL)

Alexandre Agustini (UNL)

Joaquim Ferreira da Silva (UNL)

Antônio Ribeiro (UNL)

Brasil

Docentes Doutorados

Renata Vieira (UNISINOS)

Rove Chishman (UNISINOS)

Vera Lúcia Strube de Lima (PUCRS)

Doutorandos Mestres

Sandro Rigo (UNISINOS)

Marco Antonio Insaurriaga Gonzalez (PUCRS)

3.7.Cronograma de execução

A. Em Portugal (Intercâmbio / vinda de investigadores a Portugal)

Tipo de Missão

Ano

Relativas ao refinamento, testes, utilização e avaliação do etiquetador e do analisador sintático para o Português e Português do Brasil

Investigador a designar –

2002 Setembro (15 dias)

Vera L. S. Lima –

2002 Dezembro (15 dias)

Relativas ao trabalhos sobre utilização de resolução correferência em pesquisa de informação e sistemas de gestão de diálogos

Viagens do Brasil à Portugal

Renata Vieira –

2003 Julho (20 dias)

Relativas aos trabalhos de construção de corpus anotado, esquemas de anotação e estudo de correferência em corpus paralelo

Viagens do Brasil à Portugal

Rove Chishman –

2003 Julho (20 dias)

Relativas aos trabalhos de construção de corpus anotado, esquemas de anotação e estudo de correferência em corpus paralelo

Viagens do Brasil à Portugal

Sandro Rigo –

2004 Julho (20 dias)

Marco Gonzales –

2004 Julho (20 dias)

Envio de um estudante para doutorado sandwiche.

Envio de um estudante brasileiro para doutorado em Portugal.

B. NO BRASIL (Intercâmbio / ida de investigadores ao Brasil)

Tipo de Missão:	Ano
Relativas ao refinamento, testes, utilização e avaliação do etiquetador e do analisador sintático para o Português e Português do Brasil	
Nuno M. Marques –	2002 Novembro (20 dias)
Alexandre Agustini –	2002 Dezembro (30 dias)
Relativas ao trabalhos sobre utilização de resolução correferência em pesquisa de informação e sistemas de gestão de diálogos	
Viagens de Portugal ao Brasil	
Paulo Quaresma –	2003 Agosto (20 dias)
Relativas aos trabalhos de construção de corpus anotado, esquemas de anotação e estudo de correferência em corpus paralelo	
Viagens de Portugal ao Brasil	
António Ribeiro –	2003 Dezembro (20 dias)
Relativas ao trabalhos sobre utilização de resolução correferência em pesquisa de informação e sistemas de gestão de diálogos	
Viagens de Portugal ao Brasil	
Irene Rodrigues –	2004 Agosto (20 dias)
Gabriel Lopes –	2004 Dezembro (20 dias)

3.8. Processos de acompanhamento

Conforme referido na descrição das metas propostas para o projecto, pretende-se incrementar a colaboação científica e, em consequência, o número de artigos elaborados em co-autoria. Este processo, poderá servir como uma medida de acompanhamento e de avaliação do processo de cooperação científica. Prevê-se a submissão de cerca de 5 artigos por ano a conferências internacionais.

Como meta, pretendemos iniciar dois doutoramentos nas áreas desta proposta, bem como quatro mestrados (dois no Brasil e dois em Portugal).

Propomos, ainda, efectuar relatórios anuais do projecto, avaliando o que aconteceu e perspectivando as cooperações futuras.

4. Outras referências

Lista de Projectos

PORTUGUESES

- ♦ FEELING: Projecto FEELING, contracto PBIC/C/TIT/1226/92. . Coordenador Gabriel Pereira Lopes. Financiado pela JNICT-Portugal. Participantes: GLINT, CRIA-UNINOVA, FCT/UNL.
- ♦ DIXIT: Projecto DIXIT: Multilingual Intentional Dialogue Systems. Contracto: 2/2.1/TIT/1670/95. Coordenador Gabriel Pereira Lopes. Financiado pela JNICT – Portugal.
- ♦ PGR: Projecto PGR-Acesso Selectivo aos Pareceres da Procuradoria Geral da República, contrato Nº LO59-P31B-02/97. Coordenador Gabriel Pereira Lopes. Financiado pela Agência de Inovação, Portugal. Participantes: DI-UE e GLINT, DI-FCT/UNL.
- ♦ CORPUS: projecto CORPUS, financiado no âmbito do programa Lusitânia da JNICT, via contrato PLUS/C/TIT/805/93. Coordenador Gabriel Pereira Lopes. Financiado pela JNICT-Portugal. Participantes: GLINT, DI-FCT/UNL.

- ♦ CESA: Projecto “*Córpore de Português Medieval, Etiquetagem e Segmentação Automática*”. Contracto PRAXIS 2/2.1/CSH/778/95, Coordenadores: Maria Francisca Xavier () e Gabriel Pereira Lopes. Financiado pela JNICT– Portugal. Participantes: Centro de Estudos Comparados, FCSH–UNL e GLINT, DI–FCT/UNL.
- ♦ ESA: Projecto Etiquetagem e Segmentação Automática de Corpore de Português Medieval. Contracto JNICT– FCSH/C/LIN/931/95. Coordenador: Maria Francisca Xavier. Financiado pela JNICT– Portugal. Participantes: Centro de Estudos Comparados, FCSH–UNL.
- ♦ TRADAUT: projecto TRADAUT–PT (contrato MLIS–4005 TRADAUT–PT 26192). Coordenador Gabriel Pereira Lopes. Financiado pela FCT, Portugal. Participantes: GLINT, DI–FCT/UNL, DL–FCSH/UNL e SYSTRAN.

BRASILEIROS

- ♦ ANACORT Anotação automática de co–referência textual. CNPq/FAPERGS 1999–2001. UNISINOS. Colaboração PUCRS, UNL.
- ♦ COMMON–REFs Computational model for reference. Cooperação Internacional ProTeM–CC–Inria 2001–2004. Participantes UNISINOS, PUCRS, INRIA–Nancy, França.
- ♦ MLXML Marcação lingüística de textos e documentos utilizando XML 2001–2002. UNISINOS.

COOPERAÇÃO INTERNACIONAL

- ♦ NALAMAS: financiado pelo CNPq no âmbito do programa PROTEM–CC, fase III, na área do processamento da linguagem natural. Agosto de 1996–Novembro de 1998. Universidades participantes: Brasil: PUCRS, UFRGS, UFSC, UNICAMP e USP; Portugal: Universidade Nova de Lisboa; Reino Unido: Universidade de Brighton; França: Universidade de Toulouse.
- ♦ FUNDAÇÕES — Sistemas Multi–Agentes e Processamento de Língua Natural: FUNDamentos e aplicaÇÕES, (Procº 423/CAPES). Participantes: Glint* DI–FCT/UNL e DI–Pontifícia Universidade Católica do Rio Grande do Sul. (1999–2001)

Outras Referências:

- [FaureNédellec1998] Faure, D., Nédellec, C. (1998): ASIUM: Learning subcategorization frames and restrictions of selection. 10th European Conference on Machine Learning, ECML98, Workshop on Text Mining.
- [Grefenstette1994] Grefenstette, G.: (1994) Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publisher.
- [PoesioVieira2000a] Vieira R. and Poesio M. (2000). An Empirically–Based System for Processing Definite Descriptions, Computational Linguistics, v. 26, n.4. 525–579.
- [PoesioVieira2000b] Vieira, R. and Poesio, M. (2000). Corpus–based processing of definite descriptions. In: Corpus–based and computational approaches to discourse anaphora (Studies in Corpus Linguistics, V. 3). Botley, S. and McEnery, A.M. (Eds.) Amsterdam. John Benjamins Publishing Co.
- [PoesioVieira2000c] Vieira, R. and Poesio M. (2000). Corpus–based development and evaluation of a system for processing definite descriptions, Proc. of the 18th International Conference on Computational Linguistics COLING, Saarbruecken, August. 899–905.
- [Vieiraetal2000] Vieira, R. et al (2000). Extração de sintagmas nominais para o processamento de co–referência. Anais do V Encontro para o processamento computacional da Língua Portuguesa escrita e falada PROPOR, 19–22 Novembro Atibaia SP.
- [Vilavicencioetal95] A. Vilavicencio, N. Marques, G. Lopes, F. Vilavicencio. *Part–of–Speech Tagging for portuguese Texts*. In Jacques Wainer e Ariadne Carvalho, editores, *Advances in Artificial Intelligence: Proceedings of the XII Brazilian Symposium on Artificial Intelligence*, Lecture Notes in Artificial Intelligence 991, paginas 323–332, Campinas, Outubro 11–13. Springer Verlag. 1995.