

Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa

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, Raquel Marchi, Rosário Silva, Rui Vilela, Susana Afonso

www.linguateca.pt
Diana.Santos@sintef.no

Neste artigo apresentamos uma panorâmica da actividade da Linguateca na criação e disponibilização de recursos e ferramentas para a língua portuguesa. Começamos por uma descrição dos objectivos e pressupostos da Linguateca e uma breve história da sua intervenção, e finalizamos com algumas considerações sobre a melhor forma de prosseguir na organização da área.

Apresentação

A Linguateca nasceu da identificação do processamento da linguagem natural aplicado ao português como uma área estratégica, por parte das autoridades portuguesas em Ciência e Tecnologia [33]. Após um recenseamento dos principais entraves ao desenvolvimento da área, ficou clara a necessidade da existência de recursos públicos e gratuitos e da conseqüente criação de uma infraestrutura que os disponibilizasse e que tentasse promover a colaboração entre os diversos actores, o que levou à criação de um Centro de Recursos distribuído para a Língua Portuguesa, mais tarde baptizado Linguateca. Veja-se [19] [43] [23] e [26] para diversos “instantâneos” da evolução deste projecto.

O modelo IRA – Informação, Recursos e Avaliação – preside ao funcionamento da Linguateca, que tem como pressupostos também a necessidade de organização por língua e não por âmbito geográfico [24], e a convicção de que agências de distribuição internacionais (tais como o LDC ou a ELRA) não poderão competir, para o português, com uma organização dos próprios falantes e investigadores, tanto a nível de controlo de qualidade como a nível de gestão de prioridades dos recursos humanos [22]. Do ponto de vista da organização, a Linguateca é constituída por vários pólos localizados em diferentes universidades e centros de investigação.

De forma a servir a comunidade que trata do processamento computacional da língua portuguesa, temos vários serviços genéricos: Distribuição de informação (através de um fórum constantemente actualizado, com bolsa de emprego, notícias e conferências na área), catálogo de publicações sobre o processamento de linguagem natural (PLN) do português [20], um extenso catálogo sobre os recursos existentes

para a língua portuguesa [46], um catálogo de projectos e actores, ferramentas e outra informação interessante. Além do catálogo de ferramentas/aplicações genérico, o pólo de Braga mantém também um catálogo de ferramentas livres e um serviço de Respostas a Perguntas (RaP) sobre a utilização das mesmas. Um sistema de procura dedicado, o Busca, tenta ajudar o utilizador do nosso portal.

A Linguateca funciona também como repositório de todos os recursos para o português que quiserem que nós distribuamos, quer para aumentar a sua visibilidade, quer para evitar o trabalho adicional de os distribuir.

Finalmente, a Linguateca tem promovido o paradigma das avaliações conjuntas [27],[13], através de encontros, uma lista dedicada, e sobretudo a própria organização de duas dessas acções, as Morfolimpíadas [17] [9] e o CLEF [14].

Descrição dos recursos

Os recursos produzidos ou disponibilizados pela Linguateca, por ordem cronológica do início do seu desenvolvimento, serão brevemente descritos, com ponteiros para documentação mais abrangente.

AC/DC: A primeira medida tomada para tornar os recursos já existentes mais acessíveis ao público em geral foi a criação do AC/DC (acesso a corpora/ disponibilização de corpora), servindo a consulta na rede a variados corpora [11], mais tarde anotados automaticamente pelo PALAVRAS [28].

Neste momento, o AC/DC serve mais de 250 milhões de palavras em português, nos registos jornalístico, literário, didáctico e correio electrónico [12] [10].

COMPARA: O maior corpus paralelo revisto, com textos em português e inglês e as suas traduções, e com um elevado número de utilizadores à escala mundial [3] [4].

O COMPARA é disponibilizado através do DISPARA [15], que oferece capacidades inovadoras de procura e tem uma interface rigorosamente paralela nas duas línguas.

CETEMPúblico e CETENFolha: Dois corpora de texto jornalístico de grandes dimensões [15] [42], separados em extractos, e integralmente disponíveis. Até à data registámos 281 pedidos do CETEMPúblico (Corpus de Extractos de Textos Electrónicos MCT/Público) e 131 do CETENFolha (Corpus de Extractos de Textos Electrónicos NILC/Folha de São Paulo), além das inúmeras consultas feitas através da rede.

Floresta Sintá(c)tica: A Floresta Sintá(c)tica [47] [48] é o primeiro “treebank” para a língua portuguesa: um banco de árvores sintacticamente analisadas pelo analisador sintáctico PALAVRAS de texto jornalístico (proveniente dos corpora CETEMPúblico e CETENFolha).

A Floresta tem duas partes: o Bosque, que corresponde a texto cuja anotação sintáctica foi revista e corrigida por linguistas, e que está em constante crescimento em termos de quantidade e qualidade, e a Floresta Virgem, que corresponde à parte

ainda não revista. O Bosque é disponibilizado integralmente em vários formatos, e existem dois sistemas de interrogação: o Águia [21] e o Corpuseye [8], um sistema de procura sobre o formato do Penn Treebank.

O desenvolvimento da Floresta Sintá(c)tica tem vindo a ser acompanhado por um esforço aturado de documentação, resultado das discussões conjuntas no seio da equipa do projecto, que pretende não só descrever a parte formal da Floresta mas também as diversas opções de análise linguística [49].

AnELL: O Anotador Electrónico LabEL-Linguatca [7] é um serviço público de anotação automática de textos, via web, que utiliza o INTEX [40], um sistema de processamento de linguagem natural, para produzir a anotação linguística, com base nos recursos linguísticos do LabEL [29]. De momento, os recursos incluem fundamentalmente léxicos (de unidades lexicais simples e multipalavra de vários tipos) e gramáticas de resolução de ambiguidades. São igualmente utilizados grafos para tratamento de números (cardinais, ordinais, romanos), combinações verbo-clítico e candidatos a multipalavra.

O AnELL oferece dois tipos de anotação: (i) totalmente automática; (ii) semi-automática (em fase de arranque), em que os resultados da anotação automática são revistos por um linguista, que elimina as análises incorrectas resultantes quer de sobregeração de informação (devida fundamentalmente ao elevado nível de granularidade dos recursos lexicais) quer de erros provenientes dos próprios recursos ou do processo de anotação.

À criação do AnELL presidiram dois desígnios principais: o desejo do LabEL e da Linguatca de oferecerem a qualquer utilizador um serviço gratuito de anotação linguística de textos; as vantagens, para o LabEL, de vir a melhorar a qualidade dos seus recursos através do retorno recebido de utilizadores com motivações e necessidades diferentes.

Corpógrafo: Na Universidade do Porto, o PoloCLUP da Linguatca tem desenvolvido pesquisa no uso de corpora especializados comparáveis para o estudo e a extração de terminologia [5] criando, para este efeito, o Corpógrafo [36] [37], um conjunto de ferramentas disponível 'online' a quem estiver interessado em pesquisar autonomamente. O Corpógrafo permite coleccionar textos em vários formatos, formar e analisar corpora, extrair terminologia e criar bases de dados terminológicas com uma variedade de campos e a possibilidade de criar relações semânticas e ontologias.

O Corpógrafo oferece ao utilizador, através de um simples interface na rede, a possibilidade de compilar e pesquisar os seus próprios corpora sem que para isso seja necessário ter conhecimentos especiais de informática. De certa forma, o Corpógrafo complementa a oferta de corpora genéricos já oferecido pela Linguatca possibilitando a construção e pesquisa em corpora pessoais e específicos, para utilizadores com interesses na área da Linguística, Tradução ou Engenharia do Conhecimento.

NATools: A consulta e uso em geral de corpora paralelos requerem que estes se encontrem de alguma forma interligados, nomeadamente o alinhamento ao nível da frase, segmento ou da palavra [2]. O NATools é um pacote desenvolvido no pólo de Braga da Linguatca que inclui um alinhador à frase e um outro à palavra.

Além dos alinhadores, o NATools contém um conjunto de ferramentas para tirar partido dos corpora alinhados, tais como: um gerador de dicionários probabilísticos consultáveis via rede, um módulo de classificação/avaliação da probabilidade de tradução de dois textos, um extractor de terminologia bilingue multi-palavra e um protótipo de uma ferramenta de tradução por exemplo (“example-based machine translation”).

TrAva e CorTA: O TrAva é uma ferramenta construída essencialmente para a criação de material de teste e foi desenvolvida como proposta inicial de avaliação conjunta para a tradução automática (TA). O TrAva permite traduzir frases do inglês para o português em quatro motores de TA disponíveis na Internet e apresenta um quadro de classificação de critérios linguísticos utilizando dois sistemas gramaticais: o sistema de etiquetagem gramatical utilizado pelo British National Corpus (BNC) para a classificação das frases em inglês e uma taxonomia baseada na sintaxe do português para a classificação dos resultados da tradução. Os resultados da avaliação de traduções desenvolvidos com o TrAva são consultáveis através do corpus CorTA, Corpus de Traduções automáticas Avaliadas [16].

CHAVE: No âmbito da recolha de informação (RI) cruzada (“cross-lingual”) preparámos uma colecção de documentos do Público de 1994 e 1995, junto com tópicos para RI, assim como um conjunto de perguntas e suas respostas para fazer avaliação conjunta de sistemas de resposta automática a perguntas [14].

WPT 03: Para colmatar a necessidade de colecções de documentos web em português para trabalhos de investigação sobre a língua portuguesa, a Linguatca, através do seu pólo XLDB e em parceria com o tumba!, motor de pesquisa para a Web portuguesa [38] disponibilizou este ano a maior recolha de documentos da web portuguesa existente, denominada WPT 03 [41].

A WPT 03 foi recolhida entre Março e Junho de 2003, e é composta por cerca de 3,5 milhões de documentos [6]. Adicionalmente, foi disponibilizado um diário (log) com registos das pesquisas no tumba!, com mais de 1.150.000 registos correspondentes a interrogações a esta base de documentos ao longo de 6 meses.

Esta recolha é um recurso importante para várias áreas de processamento de linguagem natural, linguística e sociologia. Além de já ser utilizada em áreas tão distintas como resposta a perguntas, procura de definições e detecção de páginas paralelas em várias línguas, pretendemos usá-la na organização da primeira avaliação conjunta em detecção e identificação de entidades mencionadas (HAREM) [31].

Esfinge: O Esfinge é um sistema de resposta a perguntas de domínio geral em português, que implementa a arquitectura descrita por Brill [30] e que explora a redundância existente na rede e o facto do português ser uma das linguagens mais utilizadas na mesma [44].

O sistema encontra-se ainda nos primórdios do seu desenvolvimento. Apesar disso, participou na tarefa Q&A do CLEF2004 [35], já que esta iniciativa era uma excelente oportunidade de avaliar o trabalho já feito, sentir algumas das dificuldades desta área e conhecer sistemas de resposta a perguntas mais evoluídos e as suas metodologias.

Memórias de tradução distribuídas: O pólo de Braga da Linguateca tem vindo a desenvolver um conceito denominado Memórias de Tradução Distribuídas (MTD), em muito semelhante à concepção das aplicações “peer-to-peer” [1]. É habitual os tradutores usarem ferramentas proprietárias de tradução que usam como memória as traduções já efectuadas, mas este processo é demasiado pessoal, já que os programas actuais não permitem a fácil partilha destas memórias.

As MTD 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. Este serviço (ainda protótipo) está implementado na tecnologia dos “Web services” e pretendemos vir a incorporá-lo em ferramentas de tradução de domínio público.

Além dos projectos descritos acima, da iniciativa total ou parcial da Linguateca, existem outros aos quais a Linguateca se associou ou funciona como parceira:

Museu da Pessoa: O Museu da Pessoa (MP) é um projecto com vista a preservar a história oral dos povos. Nasceu no Brasil, e depressa se expandiu para um núcleo em Portugal [32]. O MP é visto como um museu virtual, em que as peças em exposição não são mais do que histórias da pessoa anónima, ou seja, de todos nós que fazemos parte da história da nossa cidade e país, mas que não teríamos de outra forma a possibilidade de contar a nossa experiência de vida.

A Linguateca associou-se ao Museu da Pessoa português para tirar partido das histórias recolhidas, por constituírem uma fonte interessantíssima de corpora orais, que podem ser usados em diferentes estudos, desde o estudo da terminologia usada em diferentes zonas do país, das diferentes formas de expressão dos vários extractos sociais, até ao estudo activo do léxico.

GREASE: O pólo de Lisboa no XLDB participa no projecto GREASE (Geographic Reasoning for Search Engines) [39], o qual investiga métodos, algoritmos e arquitecturas informáticas para que um sistema auxilie o utilizador a encontrar páginas na rede, escritas em língua portuguesa, com um âmbito geográfico próximo à sua localização.

Nesse sentido, o projecto faz uso de recursos linguísticos tanto para auxiliar a fase de identificação de nomes geográficos quanto na fase de extracção de informação.

Linguarudo: Estudo da arquitectura de RI para a Web usando PLN em português, com vista à apresentação dos resultados conforme as necessidades do usuário [45]. Levará à disponibilização de um corpus de páginas da Web brasileira categorizadas por tipo de necessidade de informação.

Considerações finais

O modelo da Linguateca é o de desenvolver recursos públicos cujos beneficiários sejam a comunidade internacional que trabalha com a língua portuguesa, além de estimular a investigação no uso desses próprios recursos e na criação de outros.

Acreditamos que a partilha de experiências e de resultados é benéfica para todos os envolvidos, e que há tanto a fazer no processamento da nossa língua que não devíamos trabalhar de costas voltadas ou competindo para as mesmas fontes de financiamento.

Uma das actividades que a Linguateca se tem esforçado por patrocinar (com trabalho) é a da avaliação conjunta, em que investigadores numa dada área se reúnem para definir medidas de desempenho e o próprio formato da avaliação. Pensamos que essa é a melhor forma de pôr os investigadores a trabalhar em conjunto e de progredir no processamento da sua língua.

Referências

1. Alberto 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).
2. Alberto Simões. "Alinhamento de corpora paralelos". In J.J. Almeida (ed.), *Corpora Paralelos, Aplicações e Algoritmos Associados (CP3A)* (Braga, Junho), Braga: Univ. Minho, pp. 71-77.
3. Ana Frankenberg-Garcia & Diana Santos. "COMPARA, um corpus paralelo de português e inglês na Web". *Cadernos de Tradução IX* (2001). Universidade Federal de Santa Catarina, Brasil, pp. 61-79.
4. Ana Frankenberg-Garcia & Diana Santos. "Introducing COMPARA, the Portuguese-English parallel translation corpus". In Federico Zanettin, Silvia Bernardini & Dominic Stewart (eds.), *Corpora in Translation Education*. St.Jerome Publishing, 2003., pp. 71-87.
5. Belinda Maia. "The pedagogical and linguistic research implications of the GC to on-line parallel and comparable corpora". In José João Almeida (ed.), *Corpora Paralelos, Aplicações e Algoritmos Associados (CP3A)* (Braga, Junho), Braga: Univ. Minho, pp. 31-32.
6. Bruno Martins & Mário J. Silva. "A Statistical Study of the Tumba! Corpus". *DI/FCUL TR* 4-4, 2004.
7. Cristina Mota & Pedro Moura. "ANELL: A Web System for Portuguese Corpora Annotation". In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso & Maria das Graças Volpe Nunes (eds.), *Computational Processing of the Portuguese Language*, 6th International Workshop (PROPOR 2003) (Faro, 26-27 June 2003), Springer Verlag, pp. 184-88.
8. CorpusEye. http://corp.hum.sdu.dk/tgrepeye_pt.html
9. Diana Santos & Anabela Barreiro. "On the problems of creating a consensual golden standard of inflected forms in Portuguese". In Maria Teresa Lino et al. (eds.), *Proceedings of LREC 2004* (Lisboa, 26-28 May 2004), pp. 483-486.
10. Diana Santos & Caroline Gasperin. "Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation". In *Proceedings of the LREC'2002* (Las Palmas, 29-31 de Maio de 2002), pp. 597-604.
11. Diana Santos & Eckhard Bick. "Providing Internet access to Portuguese corpora: the AC/DC project". In Maria Gavrilidou et al. (ed.), *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)* (Athens, 31 May-2 June 2000), pp. 205-210.
12. Diana Santos & Luís Sarmento. "O projecto AC/DC: acesso a corpora/disponibilização de corpora". In A. Mendes & T. Freitas (eds.), *Actas do XVIII Encontro da Associação Portuguesa de Linguística (APL 2002)* (Porto, 2-4 Outubro 2002), APL, pp. 705-717.

13. Diana Santos & Paulo Rocha. "AvalON: uma iniciativa de avaliação conjunta para o português". In A. Mendes & T. Freitas (orgs.), Actas do XVIII Encontro da Associação Portuguesa de Linguística (APL 2002) (Porto, 2-4 Outubro 2002), APL, pp. 693-704.
14. Diana Santos & Paulo Rocha. "CHAVE: topics and questions on the Portuguese participation in CLEF". In Carol Peters (ed.), CLEF 2004 Workshop (Bath, 15-17 Setembro).
15. Diana Santos & Paulo Rocha. "Evaluating CETEMPúblico, a free resource for Portuguese". In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (Toulouse, 9-11 July 2001), pp. 442-449.
16. Diana Santos, Belinda Maia & Luís Sarmento. "Gathering empirical data to evaluate MT from English to Portuguese". In Lambros Kraniias et al. (eds.), Proceedings of the LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora (Lisboa, 25 May 2004), pp. 14-17.
17. Diana Santos, Luís Costa & Paulo Rocha. "Cooperatively evaluating Portuguese morphology". In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso & Maria das Graças Volpe Nunes (eds.), Computational Processing of the Portuguese Language, 6th International Workshop (PROPOR 2003) (Faro, 26-27 June 2003), Springer Verlag, pp. 259-266.
18. Diana Santos. "DISPARA, a system for distributing parallel corpora on the Web". In Nuno Mamede & Elisabete Ranchhod (eds.), Portugal for Natural Language Processing (PortAL 2002) (Faro, Portugal, 23-26 June 2002), Springer-Verlag, pp. 209-218.
19. Diana Santos. "O projecto Processamento Computacional do Português: Balanço e perspectivas". In Maria das Graças Volpe Nunes (ed.), V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000) (Atibaia, SP, 19 a 22 novembro de 2000), São Paulo: ICMC/USP, pp. 105-113.
20. Diana Santos. "Processamento de linguagem natural através das aplicações". In Elisabete Ranchhod (ed.), Tratamento das Línguas por Computador. Uma introdução à linguística computacional e suas aplicações. Lisboa: Caminho, 2001, pp. 229-259.
21. Diana Santos. "Timber! Issues in treebank building and use". In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso & Maria das Graças Volpe Nunes (eds.), Computational Processing of the Portuguese Language, 6th International Workshop (PROPOR 2003) (Faro, 26-27 June 2003), Springer Verlag, pp. 151-158.
22. Diana Santos. "Towards language-specific applications". Machine Translation 14, Vol.2 (1999) Dordrecht: Kluwer Academic Publishing, pp 83-112.
23. Diana Santos. "Um centro de recursos para o processamento computacional do português". DataGramZero - Revista de Ciência da Informação 3.1 (2001), fev/02.
24. Diana Santos. Porquê processamento computacional do português e não processamento de linguagem natural?. 24 de Março de 1999.
25. Diana Santos. Processamento computacional da língua portuguesa: Documento de trabalho. Livro Branco do Desenvolvimento Científico e Tecnológico Português (1999-2006), Observatório das Ciências e das Tecnologias, Ministério da Ciência e da Tecnologia, 1999.
26. Diana Santos. Relatório Linateca 2000-2003, Setembro 2003.
27. Diana Santos (ed.). Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. No prelo.
28. Eckhard Bick. The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press. 2000.
29. Elisabete Marques Ranchhod, Paula Carvalho, Cristina Mota e Anabela Barreiro. "Portuguese Large-scale Language Resources for NLP Applications", in Maria Teresa Lino et al. (eds.), Proceedings of LREC 2004 (Lisboa, 26-28 May 2004), pp. 1755-1758.
30. Eric Brill. "Processing Natural Language without Natural Language Processing", in A. Gelbukh (ed.), CICLing 2003, LNCS 2588, Springer-Verlag Berlin Heidelberg, 2003, pp. 360-369.

- 31.HAREM. <http://poloxldb.linguateca.pt/harem.php>
- 32.J. Almeida, Gustavo Rocha, Pedro Henriques, Sónia Moreira & Alberto Simões, "Museu da Pessoa: Arquitectura", Encontro da Associação de Bibliotecários e Arquivistas, 2000.
- 33.Livro Verde para a Sociedade da Informação em Portugal, Missão para a Sociedade de Informação, 1997.
- 35.Luís Costa. "First evaluation of Esfinge - a question-answering system for Portuguese". In Carol Peters (ed.), CLEF 2004 Workshop (Bath - Inglaterra, 15-17 Setembro).
- 36.Luís Sarmiento & Belinda Maia. "Gestor de corpora - Um ambiente Web integrado para Linguística baseada em Corpora". In José João Almeida (ed.), Corpora Paralelos, Aplicações e Algoritmos Associados (CP3A) (Braga, Junho), Braga: Universidade do Minho, pp. 25-30.
- 37.Luís Sarmiento, Belinda Maia & Diana Santos. "The Corpógrafo - a Web-based environment for corpora research". In Maria Teresa Lino et al. (eds.), Proceedings of LREC 2004 (Lisboa, 26-28 May 2004), pp. 449-452.
- 38.Mário Silva. "The Case for a Portuguese Web Search Engine", Proceedings of the IADIS International Conference WWW/Internet 2003, ICWI 2003, (Algarve, Portugal, 5-8 Novembro, 2003, IADIS, pp. 411-418.
- 39.Mário J. Silva, Bruno Martins, Marcirio Chaves, Nuno Cardoso, Ana Paula Afonso. "Adding Geographic Scopes to Web Resources". ACM SIGIR Workshop on Geographic Information Retrieval, Sheffield, UK, June 2004.
- 40.Max Silberstein (1993). Dictionnaires électroniques et analyse lexicale du français. Le système INTEX, Paris, Masson, 1993.
- 41.Nuno Cardoso, Mario J. Silva & Miguel Costa. "WPT 03 Recolha da Web portuguesa". In [27].
- 42.Paulo Alexandre Rocha & Diana Santos. "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa". In Maria das Graças Volpe Nunes (ed.), V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000) (Atibaia, SP, 19 a 22 novembro de 2000), São Paulo: ICMC/USP, pp. 131-140.
- 43.Pedro Veiga & Diana Santos. "Contributo para o processamento computacional do português: o CRdLP". In Maria Helena Mira Mateus (ed.), Mais Línguas, Mais Europa: celebrar a diversidade linguística e cultural da Europa. Lisboa: Colibri, 2001, pp. 103-109.
- 44.Rachel Aires & Diana Santos. "Measuring the Web in Portuguese". Brian Matthews, Bob Hopgood & Michael Wilson (eds.), *Euroweb 2002 conference* (Oxford, UK, 17-18 December 2002), pp.198-9.
- 45.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 et al. (eds.), Proceedings of LREC 2004 (Lisboa, 26-28 May 2004), pp. 1943-1946.
- 46.Signe Oksefjell & Diana Santos. "Breve panorâmica dos recursos de português mencionados na Web". In Vera Lúcia Strube de Lima (ed.), III Encontro para o Processamento Computacional do Português Escrito e Falado (PROPOR'98) (Porto Alegre, RS, 3 e 4 novembro de 1998), pp. 38-47.
- 47.Susana Afonso, Eckhard Bick, Renato Haber & Diana Santos. "Floresta sintá(c)tica: um treebank para o português". In A. Gonçalves & C.N. Correia (eds.), Actas do XVII Encontro da Associação Portuguesa de Linguística (APL 2001) (Lisboa, 2-4 Outubro 2001), APL, 2002, pp. 533-545.
- 48.Susana Afonso, Eckhard Bick, Renato Haber & Diana Santos. "Floresta sintá(c)tica: a treebank for Portuguese". In M. Rodríguez et al., Proceedings of the LREC'2002 (Las Palmas, 29-31 de Maio de 2002), pp.1698-1703.
- 49.Susana Afonso. Árvores deitadas: Descrição do formato e descrição das opções de análise na Floresta Sintáctica. Texto produzido no âmbito da Floresta Sintá(c)tica. 2004.