A Linguateca e o projecto "Processamento computacional do português"

1. Introdução

A Linguateca [7] [19] [21] [37] é um centro de recursos para o processamento computacional da língua portuguesa, em todas as variantes. Neste artigo apresentamos o trabalho já feito, a nossa actividade presente e as nossa expectativas de futuro.

Começamos por fornecer alguma informação de contexto: A Linguateca é a continuação natural do projecto "Processamento computacional do português" [17], que decorreu no SINTEF, em Oslo, de Maio de 1998 a Maio de 2000, dado que a área da engenharia da linguagem do português, considerada prioritária pelo Ministério da Ciência e da Tecnologia (MCT) português da altura [26], sofria de uma patente falta de planeamento. Nesse sentido, este projecto surgiu como a forma de o MCT promover o planeamento e reforçar o processamento da língua portuguesa, dando-lhe lugar no Livro Branco em Ciência e Tecnologia [25] e nos debates públicos sobre política científica que o precederam [18].

A intervenção da Linguateca baseia-se no modelo IRA – Informação, Recursos e Avaliação. De facto, foi por esta ordem que as actividades da Linguateca (e do projecto que a precedeu) foram evoluindo: primeiro, fez-se um levantamento do que existia e criou-se uma forma de divulgar essa informação, o portal www.linguateca.pt; depois, partiu-se para a criação dos recursos considerados mais prioritários, levando em consideração o referido levantamento; finalmente, o trabalho voltou-se para a avaliação dos recursos já existentes na comunidade, não só para tornar possível medir a evolução no futuro, como para obter indicadores sobre o grau de maturidade de cada sub-área.

Ao longo do tempo, a Linguateca foi estabelecendo cooperação com diversas instituições, nalguns casos criando pólos no interior das mesmas. O objectivo destes pólos é tornar o trabalho efectuado nestas instituições mais visível, fornecendo a infraestrutura de forma a partilhar os resultados com a comunidade em geral. Foi nesse sentido que a Linguateca se tornou numa organização distribuída, abrindo pólos 1) em Braga, no Departamento de Informática da Universidade do Minho, 2) em Lisboa, no LabEL, Centro de Automática da Universidade Técnica de Lisboa do Instituto Superior Técnico, 3) em Odense, no projecto VISL, na Universidade da Dinamarca do Sul), 4) em Lisboa, dando origem ao projecto COMPARA, 5) no Porto, no Centro de Linguística da Faculdade de Letras da Universidade do Porto, e ainda 6) em Lisboa no grupo XLDB da Faculdade de Ciências da Universidade de Lisboa, e tendo colaboração estreita com vários centros brasileiros de linguística computacional como por exemplo o NILC[1].

Há alguns pressupostos fundamentais da Linguateca que convém realçar: em primeiro lugar, a nossa matéria prima é a língua portuguesa, independentemente da variante, e por isso tentamos produzir serviços e recursos abrangendo todas as comunidades que falam português. Em segundo lugar, a nossa actividade destina-se a servir a comunidade com um todo, e daí a insistência na criação de recursos publicamente disponíveis, quer para universidades e centros de investigação, quer para empresas. Finalmente, mantemos que é essencial que sejam os falantes do português a tomarem em ombros a tarefa de avançar a área, em vez de se subordinarem ao modelo do inglês [20].

[1] http://www.nilc.icmc.usp.br

2. Informação

Um dos requisitos mais importantes para um bom planeamento consiste em saber qual o trabalho já feito. Foi nesse sentido que o projecto começou por fazer um levantamento do que já tinha sido feito na área da engenharia da linguagem do português, bem como dos recursos que poderiam ser utilizados por quem trabalhe nessa área [39].

Este levantamento permitiu estabelecer prioridades, e decidir o que era mais premente para o progresso da área. Permitiu também chegar à conclusão de que havia grande duplicação de esforços: diferentes grupos trabalhavam nos mesmos assuntos sem conhecerem o trabalho uns dos outros; havia uma falta de divulgação notória do trabalho efectuado na área. Construímos assim um portal na rede (Internet) tentando cobrir toda a actividade na área, além de fornecer informações várias de utilidade ao trabalho no processamento do português.

A manutenção deste portal é uma das tarefas que a Linguateca tem desempenhado desde a sua criação, produzindo um catálogo actualizado de recursos, ferramentas, actores, publicações e outras informações interessantes na esfera do processamento do português. Mantemos também um fórum onde divulgamos notícias, oportunidades de emprego, conferências e cursos na área, assim como a equipa da Linguateca responde a sugestões e comentários, esclarece dúvidas e tenta dar apoio a todos os membros da comunidade.

Na tabela 1 podemos ver que o interesse pelo nosso portal tem vindo a crescer como aliás o conteúdo servido pelo mesmo. A quebra observável no ano de 2003 foi provocada, tudo leva a crer, pela mudança de endereço (URL) no final de 2002.

Ano	Nº de visitas ao portal
2004 (11 primeiros meses)	657.130
2003	315.917
2002	406.298
2001	239.224
2000	124.966
1999	60.658
1998 (segundo semestre)	3.185
Total	**1.807.378**

Tabela 1: Visitas ao nosso portal

Na tabela 2 verifica-se, como seria natural, que as visitas para as quais conseguimos determinar a origem são, na sua grande maioria, efectuadas do Brasil ou de Portugal, onde se encontram maioritariamente as pessoas que trabalham em engenharia da linguagem em português.

Origem	Nº de visitas ao portal
Brasil	426.062
Portugal	211.919
Espanha	16.524
Grã-Bretanha	12.986
Alemanha	11.085
Outros países europeus	56.903
Estados Unidos	17.667
América (sem Brasil e Estados Unidos)	18.485

África, Ásia e Oceânia	13.906
Indeterminado	1.021.841
Total	**1.807.378**

Tabela 2

3. Recursos

3.1 AC/DC

O serviço AC/DC [10] [12] é o resultado da constatação de que fazia falta uma forma expedita de aceder aos corpora existentes, mesmo que fossem de livre acesso. Além de os juntar num único ponto de acesso com uma interface mais uniforme, o que foi um passo em frente, os corpora acessíveis a partir do serviço AC/DC foram ainda enriquecidos com:

a) segmentação em frases, parágrafos e unidades textuais
b) classificação gramatical e análise sintáctica fornecidas pelo analisador sintáctico PALAVRAS [22] de Eckhard Bick

Neste momento o AC/DC permite fazer pesquisas em corpora de textos jornalísticos, literários, didácticos e de correio electrónico num total de mais de 250 milhões de palavras em português.

3.2 COMPARA

O inglês é sem dúvida nenhuma o idioma mais traduzido para português. O serviço COMPARA [2] [3], lançado em colaboração com Ana Frankenberg-Garcia, dá acesso a um corpus paralelo de traduções de português para inglês e vice-versa. Oferece um grande número de opções de procura e uma interface amigável para diferentes tipos de utilizadores com diferentes necessidades de informação contrastiva.

O COMPARA é sem dúvida o maior corpus paralelo editado e revisto contendo o português. Neste momento (versão 6.0), inclui 53 textos originais e 56 traduções, e está em constante crescimento, tendo mais de uma dezena de textos em lista de espera para serem adicionados ao corpus, e tendo sido recentemente iniciada a sua anotação morfossintáctica.

3.3 CETEMPúblico e CETENFolha

Embora os projectos AC/DC e COMPARA permitissem aos investigadores estudar a língua quantitativamente e extrair várias informações interessantes, não era possível distribuí-los na íntegra, o que dificultava, por exemplo, o teste de programas sobre o material neles contido. Por outro lado, sabia-se que a existência de corpora de referência de grande dimensão em português seria bastante útil. Os corpus CETEMPúblico e CETENFolha foram uma tentativa de responder a essa lacuna.

O CETEMPúblico (Corpus de Extractos de Textos Electrónicos MCT/Público) [14] [35] contém aproximadamente 180 milhões de palavras em português de Portugal. Foi criado pelo projecto Processamento computacional do português após a assinatura de um protocolo entre o Ministério da Ciência e da Tecnologia (MCT) português e o jornal PÚBLICO em Abril de 2000. O CETENFolha (Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo) contém cerca de 24 milhões de palavras em português brasileiro, com base nos textos do jornal Folha de S. Paulo que integravam o corpus NILC/São Carlos, compilado pelo Núcleo Interinstitucional de Lingüística Computacional (NILC).

Além de serem acessíveis através do AC/DC, estando portanto segmentados em frases, parágrafos e unidades textuais e classificados gramaticalmente e sintacticamente pelo analisador sintáctico PALAVRAS, tanto o CETEMPúblico como o CETENFolha, podem ser obtidos na íntegra de um dos nossos servidores, bastando para isso que o utilizador se registe.

3.4 Floresta Sintá(c)tica

Outra das lacunas detectadas ao inventariar a área do processamento do português, foi a inexistência, para a nossa língua, de um "treebank", ou seja, de um conjunto significativo de árvores sintacticamente analisadas que permitem estudos sintácticos e avaliação de analisadores automáticos. Com a Floresta Sintá(c)tica [40] [41], lançada em 1999 em colaboração com Eckhard Bick e o projecto VISL[2], tentámos suprir essa lacuna.

Este projecto de grande envergadura conseguiu já coligir 7.756 árvores, analisadas pelo PALAVRAS e posteriormente revistas por linguistas, correspondendo a aproximadamente 150 mil palavras (versão 6.2).

3.5 AnELL

O AnELL [6] foi o resultado da conjugação das vontades, tanto do LabEL[3], como da Linguateca, em fornecer um serviço gratuito de anotação linguística de textos, em casos em que os utilizadores não pudessem ceder esses textos para consulta pública. Este serviço, acessível na rede, utiliza o INTEX [33], um sistema de desenvolvimento de sistemas de processamento de linguagem natural, para produzir a anotação linguística, com base nos recursos linguísticos do LabEL [23].

Oferece dois tipos de anotação: totalmente automática ou semi-automática. Nesta última, em fase de arranque, os resultados da análise automática são (parcialmente) revistos por um linguista.

3.6 Corpógrafo

O Corpógrafo [28] [29], desenvolvido no pólo do Porto da Linguateca permite, através de uma interface simples na rede, compilar e pesquisar corpora especializados (muitas vezes pertença exclusiva de um único utilizador) sem exigir a este conhecimentos avançados de informática.

Fornece um ambiente de trabalho que pretende resolver os problemas práticos das pessoas, dos mais básicos aos mais complexos. Por exemplo, por um lado, o Corpógrafo tem ferramentas (independentes) que extraem texto de PDF, por outro, também tenta semiautomaticamente extrair definições. Uma das suas principais funcionalidades é a extracção semi-automática de terminologia, permitindo também criar bases de dados terminológicas.

O Corpógrafo tem um artigo independente nesta edição da revista Terminómetro [4].

3.7 NATools

O NATools é um conjunto de programas desenvolvidos no pólo de Braga da Linguateca para alinhamento – ou seja, interligação de corpora paralelos [1]. O NATools inclui, além de um alinhador de frases e outro de palavras/termos, também um conjunto de ferramentas para trabalhar com corpora alinhados: um gerador de dicionários probabilísticos acessíveis pela 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 automática baseada em exemplos ("example-based machine translation").

---

[2] http://visl.sdu.dk/
[3] http://label.ist.utl.pt

3.8 TrAva e CorTA: avaliação de tradução automática para português

O TrAva (Traduz e Avalia) foi desenvolvido no contexto de uma proposta exploratória de avaliação conjunta para a tradução automática (TA) feita pela Linguateca. Esta ferramenta, desenvolvida no pólo do Porto da Linguateca, permite traduzir frases do inglês para o português em quatro motores de TA disponíveis livremente na Internet, e pede aos utilizadores para classificarem as traduções obtidas, utilizando um quadro de classificação que recorre a dois sistemas gramaticais: para as frases em inglês, o sistema de anotação gramatical utilizado pelo British National Corpus[4]; para as frases em português, uma taxonomia baseada na sintaxe do português [30]. O METRA, o sistema meta-tradutor que envia a frase em inglês para quatro sistemas distintos na rede, e que também pode ser utilizado independentemente, recebe de momento 100 pedidos de tradução por dia.

As avaliações das traduções, efectuadas pelos utilizadores do TrAva, foram armazenadas num corpus especializado, o CorTA (Corpus de Traduções automáticas Avaliadas) [9], que permite pesquisar e consultar essas traduções.

3.9 CHAVE

A colecção CHAVE [15] é um dos resultados visíveis da participação da Linguateca na organização em 2004, do CLEF (Cross-Language Evaluation Forum, Forum de avaliação conjunta cruzada) [32].

Esta colecção, um recurso útil para investigadores trabalhando na área da recolha de informação (RI), contém os textos completos do jornal diário português PÚBLICO, de 1994 e 1995, bem como uma lista de cinquenta tópicos em português, compilados em cooperação com os restantes organizadores do CLEF; as avaliações (binárias) de cada tópico, ou seja, que documentos são acerca desse 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 199 dessas perguntas. Está acessível gratuitamente dos nossos servidores, mediante um registo prévio [36].

3.10 WPT 03

A colecção WPT 03 [34] é a maior recolha de documentos da web portuguesa existente. Pode ser utilizada como um recurso importante para trabalhos de investigação em várias áreas do processamento da língua portuguesa, linguística e sociologia.

É o resultado de uma parceria entre a Linguateca e o XLDB[5] que desenvolveu o motor de pesquisa para a Web portuguesa *tumba!* [31]. Recolhida entre Março e Junho de 2003, contém aproximadamente 3,5 milhões de documentos [5]. Em conjunto com a colecção é também disponibilizado o diário (log) com os registos das pesquisas efectuadas no *tumba!* ao longo de seis meses, diário esse com mais de um milhão de registos.

3.11 Esfinge

O Esfinge [27] é um sistema de resposta automática a perguntas de domínio geral em português que explora a redundância existente na rede, bem como o facto do português ser uma das linguagens mais utilizadas na mesma [38]. O Esfinge é baseado na arquitectura proposta por Brill para o inglês [24].

Este sistema, ainda incipiente, está disponível na rede, onde é possível colocar perguntas em português e obter as dez respostas mais prováveis encontradas pelo sistema.

---

[4] http://www.natcorp.ox.ac.uk/
[5] http://xldb.fc.ul.pt/

4. Avaliação

4.1 Processo de dinamização

O primeiro passo tomado pela Linguateca na área da avaliação [13], foi estudar o que já tinha sido feito para as outras línguas nesta área. Dessa forma colheram-se ensinamentos preciosos de forma a evitar alguns dos erros cometidos no passado. Optou-se pela adopção do paradigma de "avaliação conjunta" (“evaluation contest”), segundo o qual os participantes na avaliação participam activamente na organização.

De seguida, partiu-se para a identificação de aplicações e recursos passíveis de avaliação. Criou-se um formulário na rede onde se pediu aos interessados, que indicassem as áreas que tinham mais interesse em avaliar. Este foi um passo essencial para definir quais as áreas em que havia mais potencial e interesse para organizar uma avaliação conjunta.

O passo seguinte foi a criação de uma lista de discussão, a “avalia”, destinada à discussão de todos os assuntos relacionados com avaliação de sistemas e recursos relacionados com o processamento do português. O objectivo desta lista é propiciar a discussão sobre a organização de avaliações no futuro, bem como discutir assuntos científicos e técnicos à volta da avaliação de sub-áreas do processamento do português.

4.2 Morfolimpíadas

As Primeiras Morfolimpíadas para o português [8] [11], organizadas pela Linguateca, foram a primeira actividade de avaliação conjunta realizada para o português, dedicada à análise morfológica. Decorreram de Março a Junho de 2003, culminando com a sessão final, no âmbito do encontro AVALON' 2003 - Encontro de Avaliação Conjunta de Sistemas de Processamento Computacional do Português (um encontro satélite do PROPOR' 2003 - VI Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada), organizado pela Linguateca no dia 28 de Junho de 2003 na Universidade do Algarve.

A nossa opção de começar por um exercício de avaliação na área dos analisadores morfológicos tem diversas justificações: a constatação, derivada das sondagens descritas na secção anterior, de que havia um considerável número de interessados nessa avaliação; o facto de os analisadores morfológicos serem um componente presente na grande maioria das aplicações que processam a língua portuguesa; finalmente, e atendendo ao facto de ser um acontecimento pioneiro e dada a relativa escassez de recursos materiais e humanos, a intuição de que seria a área mais realista para empreeender a primeira avaliação.

O sistema PALMORF, desenvolvido por Eckhard Bick, obteve os melhores resultados, considerando os critérios previamente estabelecidos. No entanto, todos os participantes tiveram um desempenho comparável, com diferenças apenas na faixa dos dez pontos percentuais.

Os dados e resultados desta primeira avaliação conjunta para o português podem ser obtidos a partir do nosso portal, constituindo um recurso único para quem quiser estudar os desafios que a morfologia portuguesa ainda apresenta, assim como investigar diferentes métricas de avaliação.

Nessa data foi também lançada a ideia de um livro que apresentasse ao público as actividades de avaliação conjunta do processamento computacional da língua portuguesa que têm sido levadas a cabo ou inspiradas pela Linguateca. Embora o primeiro acontecimento deste tipo, as Morfolimpíadas, ocupe uma parte significativa do livro, muitas outras áreas são focadas, visto que o objectivo principal do livro não é o de relatar simplesmente uma experiência, mas sim servir de referência a este paradigma da engenharia da linguagem em português [16].

4.3 HAREM

Além do CLEF, já mencionado acima a propósito da colecção CHAVE, a Linguateca promove neste momento o HAREM - Avaliação conjunta de sistemas de Reconhecimento de Entidades Mencionadas ("named entity recognition"). A chamada à participação ocorreu em Setembro de 2004, e contamos com que a avaliação dos sistemas participantes tenha lugar no final de Janeiro de 2005.

Escolhemos esta área para continuar o nosso esforço de avaliação em processamento da língua portuguesa por várias razões: por um lado já existe uma vasta história a nível internacional, e foi grande o interesse suscitado por uma experiência preliminar no princípio de 2003. Além disso, o HAREM permite avaliar outro tipo de capacidades das que foram objecto de estudo nas Morfolimpíadas e no CLEF: Por um lado, a tarefa é mais semântica, por outro, é menos complexa do que a resposta automática a perguntas ou a própria identificação do tópico de um documento.

5. O futuro

Pensamos que a Linguateca, nestes quase cinco anos de existência, mudou efectivamente as condições de trabalho de quem se dedica ao estudo da língua portuguesa e ao desenvolvimento de sistemas que a processam. Chegou a altura de, tendo erguido a infraestrutura, nos podermos dedicar a questões mais do foro da investigação (aplicada). Além do modelo IRA, que continuaremos naturalmente a seguir, planeamos debruçar-nos sobre os seguintes problemas: categorização automática de texto em português; extracção inteligente de terminologia bilingue sobre corpora comparáveis; investigação de padrões de uso de serviços na rede; construção semiautomática de ontologias baseadas em linguagem natural; e desenvolvimento sistemático de estruturas de indexação e procura baseadas num trabalho terminológico de base.

Agradecimentos

O agradecimento é devido a todos os que colaboram e colaboraram com a Linguateca, visto que o trabalho aqui descrito é o resultado da conjugação dos seus esforços. Adicionalmente agradecemos a tradução do resumo do artigo para espanhol e francês a Paulo Rocha e Isabel Marcelino respectivamente. A Linguateca é financiada pela Fundação para a Ciência e Tecnologia (FCT) através do projecto POSI/PLP/43931/2001, co-financiada pelo POSI.

Referências

1. 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.
2. 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.
3. 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.
4. Belinda Maia, Luís Sarmento e Diana Santos. "O Corpógrafo". Este volume.
5. Bruno Martins & Mário J. Silva. "A Statistical Study of the Tumba! Corpus". DI/FCUL TR 4-4, 2004.
6. 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.
7. 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 et al. (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.
8. 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.
9. Diana Santos, Belinda Maia & Luís Sarmento. "Gathering empirical data to evaluate MT from English to Portuguese". In Lambros Kranias 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.
10. 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.
11. 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.
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 Amália Mendes & Tiago Freitas (orgs.), Actas do XVIII Encontro da Associação Portuguesa de Linguística (APL 2002) (Porto, 2-4 Outubro 2002), Lisboa: APL, pp. 693-704.
14. 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.
15. Diana Santos & Paulo Rocha. "The key to the first CLEF with Portuguese: topics, questions and answers in CHAVE". In Carol Peters et al. (eds.), Fifth Workshop of the Cross--Language Evaluation Forum (CLEF 2004), LNCS, Springer, Heidelberg, Germany. No prelo.
16. Diana Santos (ed.). Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. No prelo.
17. 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.
18. Diana Santos. Processamento computacional da língua portuguesa: Documento de trabalho. Versão base de 9 de Fevereiro de 1999; revista a 13 de Abril de 1999.
19. Diana Santos. Relatório Linguateca 2000-2003, Setembro 2003.
20. Diana Santos. "Toward Language-specific Applications", Machine Translation 14 (2), June 1999, pp.83-112.
21. Diana Santos. "Um centro de recursos para o processamento computacional do português". DataGramaZero - Revista de Ciência da Informação 3.1 (2001), fev/02.
22. Eckhard Bick. The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press. 2000.
23. 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.
24. 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.
25. Livro Branco (1999). Livro Branco do Desenvolvimento Científico e Tecnológico Português (1999-2006), Observatório das Ciências e das Tecnologias, Ministério da Ciência e da Tecnologia.
26. Livro Verde para a Sociedade da Informação em Portugal, Missão para a Sociedade de Informação, 1997.
27. Luís Costa. "First Evaluation of Esfinge – a Question Answering System for Portuguese". In Carol Peters et al. (eds.), Fifth Workshop of the Cross--Language Evaluation Forum (CLEF 2004), LNCS, Springer, Heidelberg, Germany. No prelo.
28. Luís Sarmento & 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 2003), Braga: Universidade do Minho, pp. 25-30.
29. Luís Sarmento, 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.
30. Luís Sarmento. "Ferramentas para experimentação, recolha e avaliação de exemplos de tradução automática". In [16].

31. 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.
32. Martin Braschler & Carol Peters. "Cross-Language Evaluation Forum: Objectives, Results, Achievements", in Information Retrieval 7, Nos. 1 / 2, January/April 2004, pp. 7-31.
33. Max Silberztein (1993). Dictionnaires électroniques et analyse lexicale du français. Le système INTEX, Paris, Masson, 1993.
34. Nuno Cardoso, Mario J. Silva & Miguel Costa. "WPT 03 Recolha da Web portuguesa". In [16].
35. 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.
36. Paulo Rocha & Diana Santos. "CLEF: Abrindo a porta à participação internacional em RI do português". In [16].
37. 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.
38. 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.
39. 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.
40. 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.
41. 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.

Luís Costa
Linguateca, pólo de Oslo, SINTEF ICT
Pb 124 Blindern, N-0314 Oslo, Noruega
Tel. +47 22 06 73 11
Fax. +47 22 06 73 50
luis.costa@sintef.no

Diana Santos
Linguateca, pólo de Oslo, SINTEF ICT
Pb 124 Blindern, N-0314 Oslo, Noruega
Tel. +47 22 06 73 12
Fax. +47 22 06 73 50
diana.santos@sintef.no