O projecto Processamento Computacional do Português: Balanço e perspectivas

Diana SANTOS
Processamento Computacional do Português
SINTEF Telecom and Informatics
Postboks 124, Blindern
NO-0314 Oslo, Noruega
Diana.Santos@informatics.sintef.no

Abstract

This paper describes the Computational Processing of Portuguese project, providing an overview of its work in three areas: the building and maintenance of a resource catalogue for NLP of Portuguese; the serving of corpora through the World Wide Web; and the evaluation of tools and resources. The paper emphasises strategic options, presenting mainly material not subjected to (previous) scientific publication, such as the administration of the Web pages, the version numbering of corpora, and the plans for tool distribution. In the second part, a distributed resource center in its creation phase is presented, which is the logical continuation of the work of the past two years.

Resumo

Após dois anos de trabalho orientado para a melhoria do panorama do processamento computacional do português, urge fazer um balanço e discutir abertamente quais as perspectivas de continuação futura. Neste artigo descrevo em traços largos a actividade passada do projecto Processamento Computacional do Português nas vertentes catálogo, processamento de corpora e avaliação, discutindo algumas questões subjacentes aos serviços que oferecemos. A ênfase é posta nos problemas que tentámos resolver, na estratégia seguida, e na discussão das alternativas. Finalizo com a minha visão do que deveria ser um centro de recursos distribuído para o processamento computacional da língua portuguesa, de momento em fase embrionária.

1. Apresentação

O projecto Processamento Computacional do Português foi lançado em Maio de 1998 como uma primeira medida para organizar a área da engenharia da linguagem do português, considerada pelo Ministério da Ciência e da Tecnologia (MCT) uma das suas prioridades em Portugal, da qual era patente, contudo, a debilidade a nível nacional e internacional.

O mandato do nosso projecto era, pois, relativamente vago, ainda que reflectisse uma consciência aguda das necessidades e das enormes carências, fossem elas de material humano, recursos materiais, ferramentas ou mesmo de um fórum que reunisse a comunidade. Igualmente evidente era a falta de formação básica.

Neste artigo pretendo apresentar a evolução da nossa actividade e as ilações que pudemos tirar. Não explicarei as razões da necessidade de investir na área nem a motivação para o lançamento do projecto nos moldes em que o foi feito, visto que essa informação já foi vastamente discutida e divulgada, quer antes do lançamento do projecto (Livro verde) quer durante as primeiras fases do mesmo (Santos, 1999a).

Este projecto foi concebido como uma fase temporária de planeamento e intervenção no processamento da língua portuguesa por parte do MCT, associado que estava à criação do Livro Branco em Ciência e Tecnologia (Livro Branco) e aos debates públicos sobre

política científica que o precederam. Uma das suas eventuais missões seria viabilizar projectos ou iniciativas de maior alcance como continuação ou resultado.

A nossa escolha foi lançar um centro de recursos – distribuído – para a língua portuguesa, proposta que foi informalmente aprovada pelo MCT em Janeiro deste ano, e formalizada com o SINTEF em Abril. Essa proposta vem assim reforçar o trabalho primordialmente centrado em recursos e na sua disponibilização a que nos temos dedicado no segundo ano da nossa actividade, e que lançou as bases, a nível de organização e de contactos científicos, do centro que pretendemos criar.

É tempo, pois, de um balanço de forma a ponderar os resultados e as fraquezas da actividade passada. E, além disso, urge partilhar com a comunidade os nossos planos de forma a obter comentários e críticas a tempo de poder inflectir e servir melhor aqueles que pretendemos servir: todos os que se dedicam ao processamento computacional da nossa língua.

Este artigo encontra-se dividido em duas partes:

1. **Passado**, em que me concentro sobre o trabalho feito, evitando contudo a mera descrição do que se encontra acessível a todos na Web. O meu objectivo é falar do trabalho subjacente, das opções não evidentes e de alguns comentários críticos que se me afiguram relevantes.
2. **Futuro**, em que além dos projectos que já iniciámos e tentaremos levar a bom termo, e que serão brevemente relatados, descreverei as linhas de força do que me parece que deveria ser um centro de recursos distribuído para a língua portuguesa e as actividades que gostaríamos de impulsionar.

Com este texto gostaria de envolver todos quantos se revêem na área e pedir comentários, sugestões e iniciativas de colaborações conjuntas, de forma a que o pouco que pudermos fazer contribua efectivamente para a melhoria da nossa comunidade.

2. Passado

Não pretendo reflectir exaustivamente sobre tudo o que fizemos nestes dois anos; para efeitos do presente artigo, debruço-me apenas sobre três vertentes da nossa actividade que gostaria de divulgar melhor e sobre as quais receber reacções críticas: o catálogo, os corpora, e a avaliação.

2.1 Catálogo: A parte mais visível

Ainda que tenhamos começado a catalogação da área na Web como uma reacção à falta de informação e comunicação que existia em Portugal, criando assim uma fonte alternativa de medição e observação da área, e permitindo uma maior conhecimento mútuo dos diversos intervenientes, cedo nos apercebemos que a manutenção de um catálogo como o nosso podia ser, por si só, um serviço para a comunidade, constituindo-se num portal para o processamento da língua portuguesa.

Como qualquer visitante de http://www.portugues.mct.pt tem oportunidade de verificar, apontamos para um número considerável de endereços relacionados com o processamento do português, ainda que uma análise mais atenta permita compreender que, em muitos casos, esses endereços simplesmente mencionam, ao invés de oferecerem (ou venderem), serviços ou recursos.

Para tentar dar uma ideia imediata do tipo de acessibilidade, indicamos, no caso dos recursos, a forma de distribuição/acesso através de um pequeno conjunto de ícones. No caso dos projectos (correspondentes a 82 endereços) não temos qualquer forma de indicar se estes deram origem a resultados concretos e qual o seu estatuto (sobretudo porque, como é sabido, as páginas da rede tendem a apodrecer rapidamente, ou seja, a deixarem de ser válidas ou até passíveis de modificação pelos seus autores). Por isso, o número por si

só de páginas listadas, além de poder reflectir a dispersão de recursos económicos e de temas tratados, terá um interesse predominantemente histórico.

Mesmo no caso dos recursos, a linearidade do nosso catálogo pode ser enganadora. De facto, na esmagadora maioria dos casos, o número de actores ou recursos distintos não é um dado suficientemente informativo: por exemplo, um grupo ou um projecto – a que poderá apenas corresponder um endereço, pode representar mais de três quartos dos recursos disponíveis, espalhado o último quarto por quinze actores diferentes. Da mesma forma, a existência de quinze conjugadores verbais para o português não significa que a sua qualidade esteja assegurada, nem que o único sintetizador de fala para o português, por ser único, seja de pouca qualidade.

Além disso, muitos dos sistemas mais complexos incluem como partes, não separadamente identificáveis, outros sistemas mais simples. Não fez, no entanto, sentido para nós, ao listar um analisador sintáctico ("parser"), também o incluir sob as entradas analisador morfológico e léxico nas categorias correspondentes.

É também evidente que muitos dos recursos não são comparáveis, no sentido do tempo necessário para os criar, da qualidade do seu funcionamento, do cuidado posto na sua documentação ou distribuição, etc.

Em suma, embora o resultado seja aparentemente útil e fácil de compreender pelos visitantes das nossas páginas, temos consciência clara das suas limitações, não obstante termos tentado minimizar o grau de subjectividade posto na criação do catálogo (Oksefjell & Santos, 1998):

- Mantendo os nomes dados pelos autores das páginas
- Listando por ordem alfabética
- Não fazendo quaisquer juízos de qualidade (por exemplo, qualquer recurso que afirme fazer tradução automática para português é introduzido no catálogo, mesmo que a qualidade dessa tradução seja francamente má) nem de adequação terminológica (qualquer lista de palavras identificada como dicionário é adicionada a esta categoria)
- Colocando o mesmo endereço sob várias categorias quando uma dada localização na rede se refere a mais de um recurso

Não deixamos, contudo, de ter consciência de que a categorização é um problema extremamente complexo e que nunca poderá ser resolvido por uma estrutura hierárquica simples. Além disso, a quantidade de informação para que apontamos começa a tornar difícil a um utilizador escolher que caminhos percorrer dentro do nosso sítio de forma a chegar às páginas que lhe interessam. De facto, é cada vez mais frequente, no nosso quotidiano, que ao deparar-se-nos uma referência a um dado recurso ou sítio de interesse, tenhamos dificuldade em confirmar, através da simples navegação pelas nossas páginas, se já se encontra no nosso catálogo.

Por estas razões criámos dois sistemas:

- o **Menuseador**, que automatiza a criação das páginas (a partir de um índice interno) de forma a permitir reformulações de classificação com simplicidade, e que é uma ferramenta interna do nosso projecto (desenvolvida pelo Paulo Rocha)
- o **Busca**, um sistema de busca sobre o conteúdo do nosso catálogo que permite a um utilizador chegar mais depressa às páginas procuradas (desenvolvido pelo Tom Funcke e adaptado pelo Paulo Rocha)

Em relação a este segundo sistema, é interessante mencionar que continua a ser muito maior o tráfego das visitas ao catálogo do que o uso do sistema de busca, o que pode significar que o agrupamento dos recursos por categoria tem a sua utilidade. (Outras explicações podem ser a habituação dos utilizadores à estrutura do catálogo – visto que o sistema de busca é relativamente recente; a sua descrença em relação a ferramentas de

procura, muitas vezes pouco cooperativas; ou então a sua preferência pelos grandes motores de procura quando têm uma pergunta específica).

Um exemplo da agilidade que o primeiro permite foi a recente adição da categoria **Recursos:Material didáctico:Cursos de literatura**, em que bastou editar o ficheiro de menus adicionando a seguinte linha:

```
3×N××Cursos de literatura×didactico.html#did3lit
```
[1]

E editar, para cada recurso, a sua nova localização:

```
a×did3lit×Panorama da Literatura Brasileira×http://www.nilc.icmsc.sc
.usp.br/literatura/bemvindo.htm××000428×pt
```
[2]

Invocando em seguida o Menuseador, as páginas com a nova estrutura foram automaticamente criadas. Convém, contudo, mencionar que este sistema permite a continuação da manutenção do catálogo editando os próprios ficheiros de HTML, que são a base de todo o catálogo, ou seja, não se complicou o processo no caso de vários colaboradores diferentes terem modos diferentes de efectuarem modificações ao catálogo (Rocha, em preparação).

Uma outra actividade que consideramos de catálogo, mas com problemas próprios, é a lista de publicações, não só pela vagueza da área em termos gerais, mas pelo facto de os títulos, quer dos próprios artigos quer dos livros ou conferências em que são publicados, não serem em muitos casos suficientemente esclarecedores para permitir avaliar com clareza se a obra deve ou não ser incluída como relacionada com o processamento computacional da nossa língua.

Finalmente, convém mencionar que, ao lado do apreço que nos parece generalizado em relação à nossa actividade catalogadora, temos tido várias críticas em relação à sua aparência. O aspecto gráfico é, contudo, algo extremamente subjectivo e sujeito mesmo a regras contraditórias conforme a comunidade de origem – o que tem levado, aliás, à própria falta de consenso no interior do projecto.

2.2 Corpora: a parte mais trabalhosa

Por ser uma área de intervenção em que a necessidade era unânime e em que não havia o perigo de ameaça a interesses comerciais (não há nenhuma empresa que venda ou invista em corpora de texto em português – ao contrário de dicionários ou ferramentas computacionais), resolvemos começar por ela, aliás também a mais simples quando se trata da sua criação (já o mesmo não se pode dizer em relação à sua exploração).

Lançámos portanto o projecto AC/DC (Acesso a Corpora/Disponibilização de Corpora), que numa primeira fase se encarregou de verter todos os corpora já existentes para um sistema de manipulação comum e dar-lhes acesso através da rede (Santos, no prelo b). Numa segunda fase encontramo-nos a proceder à análise automática desses mesmos corpora, de forma a permitir procuras muito mais elaboradas (Santos & Bick, 2000).

Tornou-se, contudo, patente a falta de material, sobretudo para o português europeu, o que fez com que nos lançássemos também na criação do CETEMPúblico, um corpus de 180 milhões de palavras de linguagem jornalística portuguesa, distribuído em CD e na rede (Rocha & Santos, 2000).

Não pretendo repetir aqui aquilo que já foi dito nos artigos acima mencionados, mas antes descrever algumas outras questões associadas ao processamento e/ou criação destes recursos.

---

[1] nível, página de menu própria (N=não), nome da página (se tiver), título da categoria, endereço, identificação interna

[2] validade (a=válido, x=suspenso, não é para incluir no catálogo), did3lit=categoria, título, endereço(s), figura(s) data, língua(s)

No decorrer do projecto AC/DC, criámos um conjunto significativo de ferramentas de processamento de corpora, desde programas de limpeza e tratamento de corpora específicos a separadores de frases genéricos. Pensamos que seria útil disponibilizar este banco de programas, que podem ser úteis mesmo no caso de os corpora não poderem ser acedidos através do nosso serviço na rede.[3] Contudo, é evidente que torná-los disponíveis exige um esforço considerável de documentação.[4]

O que é, aliás, um problema geral com a questão dos corpora.

Falta de documentação a vários níveis: como usar um corpus? Quais os critérios postos na sua compilação, quais os problemas – resolvidos ou não – a ele associados? O que é que é necessário para a sua caracterização (mínima)[5]? Devemos investir na informação nas páginas da Internet/relatórios técnicos, ou em artigos em conferências internacionais? Em relação ao CETEMPúblico, e visto que foi criado por nós, pareceu-nos evidente a segunda opção. Mas quando disponibilizamos um corpus que não tem documentação (suficiente), temos de ser nós a produzi-la? O que fazer quando temos versões dos corpora que diferem – por variadas razões – das existentes nas mãos dos seus compiladores? Ou quando estes variam a codificação ou conteúdo do corpus sem se preocuparem com a identificação das versões?[6]

Além disso, é preciso lembrar que a disponibilização dos corpora como nós a efectuamos é um processo dinâmico: de cada vez que depuramos os programas (e fazemo-lo sempre que descobrimos problemas não tratados anteriormente), temos uma nova versão da codificação do corpus. (Assim, é preciso fazer a gestão das versões do corpus codificado, e do corpus "cru").

A questão das versões é problemática ainda de um terceiro ponto de vista, e que é o de manter a compatibilidade com os investigadores que poderão usar os mesmos corpora, sem ser através do nosso serviço. Se nós limparmos ou alterarmos um corpus, o mesmo acontecendo com outros grupos, a probabilidade de obtermos objectos diferentes é extremamente elevada. Por essa razão, decidimos nunca alterar o conteúdo original, e criar programas de limpeza que podem ser distribuídos por outros grupos mas que partem sempre do objecto inicial. Do ponto de vista de processamento, é pesado, mas pensamos que vale a pena.

Outra questão que gostava de levantar é a da necessidade de haver um corpus de referência que possa ser usado por todos como uma medida de comparação, e que possa ser calibrado de forma a comparar o desempenho de diversos grupos em diversas tarefas. Isso foi uma das razões que nos levou a criar o CETEMPúblico da forma que o fizemos, não obstante haver já vários grupos em Portugal e no estrangeiro que armazenam semanalmente, com ou sem autorização, as edições electrónicas deste jornal. Em primeiro lugar, a partir de agora todos o podem usar sem repetir o mesmo trabalho de recolha; em segundo lugar, podem comparar resultados com base no mesmo conjunto, e com uma numeração/identificação padrão (em termos de extractos), porque distribuída com o

[3] Por exemplo, evitaria duplicações de esforços na preparação da parte portuguesa do corpus MLCC (Armstrong et al., 1998), que é possível comprar à ELDA, mas não distribuir a outrem. Além disso, permitiria diminuir o esforço posto na preparação de corpora locais com um formato parecido com os já processados por nós.

[4] Interessa-nos, por isso, averiguar o interesse da comunidade em ter acesso a esses programas e qual a documentação de que necessitariam.

[5] De momento, temos automatizada a obtenção de algumas características tais como o número de unidades, o número de palavras, o número de frases, o número de parágrafos, etc., além de fornecer a instituição de origem, os endereços da rede e a bibliografia que conhecemos. Muitas outras características seriam, evidentemente, relevantes.

[6] É evidente que lhes cabe esse direito. Mas devemos nós tentar seguir as alterações de perto e numerar as versões cruas, ou devemos pura e simplesmente fixar-nos numa versão com uma dada data e não mudar o corpus cru, aconteça o que acontecer?

corpus.[7] Além disso, separámos o texto em frases e parágrafos, sinalizando títulos e autores. Pensamos, assim, aumentar o valor do corpus, ainda que introduzamos certamente alguns erros. Aos utilizadores cabe a opção de fazer ou não uso desta informação, que é, aliás, trivial ignorar.

Em terceiro e último lugar, não é despropositado falar aqui de formação. Por um lado, sou de opinião que fornecer corpora e mais nada assemelha-se a dar um carro a quem não sabe guiar, ou doar um edifício para uma biblioteca, mas não livros. A existência dos corpora é necessária, mas não suficiente, para se avançar no processamento da nossa língua. Por outro lado, a oferta de um carro pode também ser considerado um incentivo para aprender a conduzir, e a de um edifício para começar a juntar livros.

Convém que reconheçamos que não é fácil utilizar corpora, para além da mera confirmação de se uma palavra se encontra atestada ou não. Tive várias vezes ocasião de demonstrar que não é trivial usar um corpus para obter conclusões em linguística (Santos, no prelo c); por outro lado, sem conhecer e poder medir as diferenças entre vários corpora (ou tipos de textos diferentes) não é possível extrapolar medidas de desempenho de sistemas treinados num dado corpus nem mesmo validar generalizações feitas com base em corpora (Santos & Oksefjell, 1999). Por outras palavras, utilizar corpora é difícil, mas é o único caminho. Contribuir para a sua existência é, por isso, contribuir para a possibilidade do surgimento de trabalhos de qualidade sobre a nossa língua – embora não automaticamente.

2.3 Avaliação: a parte mais difícil

Outra das actividades a que o nosso projecto se dedica é a avaliação. Efectuámos o estudo de um alinhador de corpora paralelos (Santos & Oksefjell, 2000), de todos os conjugadores verbais a que tivemos acesso (Rocha, 2000), assim como a comparação de duas ferramentas de corpora utilizadas para nossa língua (Santos & Ranchhod, 1999). De uma forma implícita, também procedemos à avaliação – ainda não documentada – de motores de busca em português e de corpora da nossa língua (veja-se uma primeira comparação destes últimos em Santos (no prelo b)).

Conforme já foi mencionado quando me referi ao catálogo, a avaliação exige um conhecimento elevado do problema e o desenvolvimento de metodologias próprias. É, pois, uma tarefa vastíssima, mas aquela que me parece requerer a maior atenção do ponto de vista do estudo e desenvolvimento da nossa área. Cito, a esse propósito, Hirschmann (1998:302, tradução minha), que afirma que "a avaliação é em si própria uma actividade de investigação de primeira classe: a criação de métodos de avaliação efectivos leva a um progresso rápido e a melhor comunicação no seio de uma comunidade científica".

3. Futuro

Começo por mencionar sucintamente o prosseguimento natural da actividade nas três áreas discutidas acima, explicando depois como pretendemos dar alma e corpo ao centro de recursos distribuído, cujos princípios, organização e objectivos serão brevemente delineados.

3.1 Catálogo

Após um início pouco prometedor do nosso serviço de repositório, em que muito poucos investigadores usaram a nossa oferta para disponibilizar os seus sistemas, publicações ou

[7] Tal não impede que os grupos de investigação, na posse de todo o material do CD, apaguem toda a anotação/identificação fornecida para efeitos das suas tarefas internas. Não há limitações à investigação feita com o conteúdo do CD; para comparar resultados, contudo, parece-nos de interesse utilizar a identificação por nós fornecida.

serviços, pensamos que devemos investir na nossa capacidade de espelho e não meramente de índice.

Tencionamos avançar no sentido da dinamização do catálogo, permitindo alguma adaptação ao utilizador, quer iniciada por este, quer baseada estatisticamente no perfil de visitas. Esta possibilidade poderá, além disso, resolver, pelo menos parcialmente, a questão da aparência.

3.2 Corpora e outros recursos

Além da óbvia continuação da criação de outros corpora a partir de material já acessível, tal como discurso literário, correio electrónico / listas de discussão, a Web como corpus, etc., encontramo-nos neste momento na fase inicial do projecto COMPARA / DISPARA, iniciado por Ana Frankenberg-Garcia e ao qual aderimos, cujo objectivo é a compilação de um corpus paralelo português-inglês e sua disponibilização na rede (Frankenberg-Garcia & Santos, no prelo).

Tencionamos também abrir caminho na facilitação do acesso e/ou na construção de outro tipo de recursos, tais como ferramentas e léxicos.

3.3 Avaliação

Parece-nos evidente que o ideal seria obter o maior número de estudos sobre vários tipos de ferramentas diferentes. A única maneira de concretizar este objectivo é encomendar – como uma das atribuições do centro – diferentes avaliações a diferentes investigadores, dando o máximo de divulgação às já existentes ou em curso.[8]

3.4 Organização e concepção do centro

Os principais objectivos do centro[9] podem ser resumidos da seguinte forma: facilitar o acesso aos recursos já existentes, desenvolver de forma harmoniosa e em colaboração com os interessados aqueles considerados mais prementes, organizar avaliações e conferências a que chamei "avaliações conjuntas" em Santos (1999b), velar por que os recursos encaminhados para esta área possam aproveitar ao máximo o progresso desta, manter o catálogo actualizado e melhorar o portal como um todo.

Além disso, parece-nos importante que o centro fomente o ensino da área através da Web, tentando também incentivar a criação de textos, sistemas pedagógicos e material de teste na área do processamento computacional do português. Consideramos, pois, a ideia de encomendar, além das avaliações mencionadas acima, lições e panorâmicas sobre diversas áreas a peritos, numa chamada pública.[10]

A actividade do centro repartir-se-á entre

- a formação de pessoal especializado em gestão de recursos
- a gestão de um programa de desenvolvimento de recursos (incluindo recursos de formação) por concurso público
- o assegurar dos serviços básicos de repositório, distribuição e catálogo, lançando as bases para tal vir a ser feito de forma distribuída
- o desenvolvimento de alguns recursos pelo próprio centro, sobretudo recursos para avaliação ou para calibragem

Duas linhas mestras nortearão a actividade do centro:

---

[8] Através de contactos directos (Luís Caldas de Oliveira e Marco Esteves da Rocha) tivemos conhecimento de avaliações de sintetizadores de fala para português europeu e de anotadores/taggers para português escrito.

[9] A palavra *centro* não é ideal, devido ao choque com o adjectivo *distribuído*. Contudo, as alternativas *organismo*, *instituição*, *instituto*, *rede*, *núcleo*, *pólo* ou *biblioteca* também não nos pareceram satisfatórias.

[10] Em Santos (no prelo a) tentei uma primeira apresentação de ferramentas / aplicações.

1. **Total abertura**: Todas as chamadas, actividade e propostas submetidas serão públicas. Apenas os pareceres a entidades oficiais portuguesas, encomendados por estas, o poderão não ser.
2. **Disponibilização livre dos recursos**: Os autores serão remunerados de forma a não serem lesados, mas este centro não se destina a desenvolver ou apoiar o desenvolvimento de recursos proprietários. Pelo contrário, destina-se a criar condições para a existência de recursos bons e grátis para a nossa língua.

O presente artigo tem a data de 1 de Agosto de 2000. Por ora, encontramo-nos na fase de instalação do embrião de um centro, regido pela Fundação para a Computação Científica Nacional (FCCN), que, além do pólo de Oslo, contará com um pólo em Lisboa para tratamento de recursos de fala, e um pólo em Braga primordialmente dedicado a ferramentas Linux. Esperamos que, daqui a um ano, já seja visível a actividade plena deste centro distribuído.

Agradecimentos

Este artigo – e o projecto como um todo – deve muito a várias pessoas. Em primeiro lugar aos outros membros do projecto: Signe Oksefjell, Paulo Alexandre Rocha e Tom Funcke, mas também a todos os investigadores que com ele directamente colaboram: Elisabete Ranchhod, Eckhard Bick e Ana Frankenberg-Garcia, ou forneceram recursos e/ou apoio diverso, entre os quais saliento: José João Dias de Almeida, Maria da Graça Nunes, Denise Kuhn, Isabel Trancoso, Luís Caldas de Oliveira e Tony Berber Sardinha.

Da mesma forma, cabe exprimir aqui a gratidão a todos quantos providenciaram informação, encorajamento e, sobretudo, recursos (ou autorização para os disponibilizarmos).

Referências

AC/DC. Acesso a Corpora / Disponibilização de Corpora http://cgi.portugues.mct.pt/acesso/

Armstrong, Susan, Masja Kempen, David McKelvie, Dominique Petitpierre, Reinhard Rapp & Henry S. Thompson (1998). Multilingual Corpora for Cooperation. In Antonio Rubio, Natividad Gallardo, Rosa Castro and Antonio Tejada (eds.), *Proceedings of The First International Conference on Language Resources and Evaluation* (Granada, 28-30 May 1998), Vol. 2, pp. 975-80.

Frankenberg-Garcia, Ana & Diana Santos (no prelo). Introducing COMPARA, the Portuguese-English parallel translation corpus. *Proceedings of The Second International Conference on Corpus Use and Learning to Translate, CULT 2K* (Bertinoro, 3-4 November 2000).

Hirschman, Lynette (1998). The evolution of Evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language* **12** (4), pp.249-62.

Livro Verde (1997). *Livro Verde para a Sociedade da Informação em Portugal*, Missão para a Sociedade de Informação, 1997, http://www.missao-si.mct.pt/livroverde/livrofin.htm.

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, http://www.mct.pt/Livro-BrancoCT/Welcome2.html.

Oksefjell, Signe & Diana Santos (1998). Breve panorâmica dos recursos de português mencionados na Web. In Vera Lúcia Strube de Lima (ed.), *Anais do Terceiro Encontro de Processamento da Língua Portuguesa (Escrita e falada), PROPOR'98* (Porto Alegre, 3-4 novembro 1998), pp. 38-47.

Rocha, Paulo Alexandre (2000). Uma apreciação de diversos recursos para conjugação de verbos em português. SINTEF, Oslo, 2 de Fevereiro de 2000, http://www.portugues.mct.pt/Paulo/pubs/conjug.html.

Rocha, Paulo Alexandre (em preparação). Gestão das páginas do projecto Processamento computacional do português. SINTEF, Oslo.

Rocha, Paulo Alexandre & Diana Santos (2000). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)* (Atibaia, São Paulo, Brasil, 19 a 22 de Novembro de 2000), este volume.

Santos, Diana (1999a). Processamento computacional da língua portuguesa: documento de trabalho. SINTEF, Oslo, versão base de 9 de Fevereiro; revista a 13 de Abril, http://www.portugues.mct.pt/branco/.

Santos, Diana (1999b). Disponibilização de corpora através da WWW. In Palmira Marrafa & Maria Antónia Mota (eds.), *Linguística Computacional: Investigação Fundamental e Aplicações*. *Actas do I Workshop sobre Linguística Computacional da Associação Portuguesa de Linguística* (Lisboa, 25-27 de Maio de 1998), APL, Lisboa: Colibri, pp.323-346.

Santos, Diana (no prelo a). Introdução ao 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, no prelo

Santos, Diana (no prelo b). Comparação de corpora em português: algumas experiências. In Tony Berber Sardinha (ed.), *Língua Portuguesa no Computador*, São Paulo.

Santos, Diana (no prelo c). Aonde vamos em relação a *aonde*. Apresentado no Simpósio *´Redescobrindo a linguagem: Pesquisa em Linguística de Corpus´, 10.º InPLA* (São Paulo, 14 de abril de 2000), a ser publicado em *Intercâmbio* **10**.

Santos, Diana & Eckhard Bick (2000). Providing Internet access to Portuguese corpora: the AC/DC project. In Maria Gavrilidou et al. (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation* (Athens, 31 May-2 June 2000), pp.205-210.

Santos, Diana & Elisabete Ranchhod (1999). Ambientes de processamento de corpora em português: Comparação entre dois sistemas. *Actas do IV Encontro sobre o Processamento Computacional da Língua Portuguesa (Escrita e Falada), PROPOR* (Évora, 20-21 de Setembro 1999), pp. 257-268.

Santos, Diana & Signe Oksefjell (1999). Using a Parallel Corpus to Validate Independent Claims. *Languages in contrast* **2(1)**, 1999, pp.117-132. John Benjamins Publishing Co.

Santos, Diana & Signe Oksefjell (2000). An evaluation of the Translation Corpus Aligner, with special reference to the language pair English-Portuguese. In Torbjørn Nordgård (ed.), *NODALIDA'99, Proceedings from the 12th "Nordisk datalingvistikkdager". Trondheim, 9-10 December 1999*. Trondheim: Department of Linguistics, NTNU, 2000, pp.191-205.