Um Centro de Recursos para o Processamento Computacional do Português

A Resource Centre for the Computational Processing of the Portuguese Language

por Diana Santos

Resumo: Neste artigo apresento o Centro de recursos para o processamento computacional da língua portuguesa e o projecto Processamento computacional do português, mencionando as razões que levaram à sua criação e expondo os seus objectivos e actividade. Após explicar o que se entende por processamento computacional de uma língua e de que recursos se trata, apresento brevemente as três vertentes de trabalho já realizado ou em curso: a disseminação da área, a disponibilização de recursos e a avaliação.
Palavras-chave: Processamento de linguagem natural; Língua portuguesa; Recursos linguísticos; Serviços na rede; Avaliação; Corpora

Abstract: In this paper the newly created Resource Centre for the Computational Processing of the Portuguese Language and the Computational Processing of Portuguese project are introduced. It starts by motivating the creation of the center, and continues by explaining its goals and main activity. The natural language processing discipline (with an emphasis on applications) is then presented, with a description of what kind of resources it requires. The main bulk of the paper is devoted to an overview of the three different kinds of activity of the centre: dissemination, resource distribution, and evaluation.
Keywords: Natural language processing; Portuguese language; Linguistic resources; Services on the net; Evaluation; Corpora

Neste artigo descrevo o trabalho realizado para a agilização do processamento computacional da nossa língua, pelo projecto "Processamento Computacional do português", mais tarde evoluindo para um Centro de Recursos distribuído.

1. Motivação

A área do processamento computacional da língua portuguesa foi uma das áreas consideradas prioritárias na política de ciência e tecnologia pelo governo português (veja-se o *Livro Verde para a Sociedade da Informação em Portugal (1997)* e o *Livro Branco do Desenvolvimento Científico e Tecnológico Português (1999-2006)*). Além disso, foi também identificada como uma área especialmente problemática, devido a vários factores, dois dos quais pretendemos (parcialmente) colmatar com o presente centro:

* a falta de recursos que possam servir de base ao desenvolvimento de aplicações e ao próprio estudo da língua portuguesa;
* e a falta de métodos e de métricas de avaliação para comparar sistemas e para avaliar o progresso na área.

Ao lançar o projecto e mais tarde o centro, o objectivo genérico foi, portanto, o de criar uma estrutura de disseminação e agilização da construção e disponibilização de recursos para o processamento (computacional) da língua portuguesa, com a qual se pudesse então proceder à avaliação do progresso na área.

Desde o início que não se pretendeu o lançamento de mais uma estrutura pesada, dirigida e dominada por burocratas, mas sim uma acção que reunisse diversos grupos espalhados em localizações diferentes, que comungassem de um espírito comum de partilha de resultados e experiências. Assim, o centro é constituído por pólos, que, ao mesmo tempo que criam novos serviços ou recursos para a comunidade, tentam a integração e potenciação do espírito e da filosofia do projecto inicial em centros de investigação de renome e experiência em processamento do português, absorvendo algum do saber-fazer desses grupos e tornando real a colaboração prática, no sentido de desenvolver e disseminar os recursos já existentes.

De momento, o centro de recursos tem três pólos em funcionamento:

* do SINTEF em Oslo, o pólo seminal
* do Departamento de Informática da Universidade do Minho em Braga
* do Laboratório de Engenharia da Linguagem do Instituto Superior Técnico em Lisboa

Outros tantos estão em fase de organização em Portugal, assim como pretendemos, através do estabelecimento de protocolos de cooperação internacional com o Brasil e outros países lusófonos, o mais breve possível estendê-lo a grupos nestes países.

O centro foi pensado como uma forma de endogeneizar certos padrões de comportamento, e normas de conduta, entre os principais actores do processamento da língua. Na proposta de constituição, foi sublinhado que se pretendia que fosse *uma estrutura temporária, mais tarde podendo vir a ser absorvida pelas próprias instituições que o*

albergam.

No presente artigo tentarei descrever o passado, o presente e o futuro da nossa actividade. Convém, contudo, ressaltar que a actividade central do centro não é de pesquisa, mas sim de acção concreta de política científica de forma a atingir os dois objectivos mencionados acima, não em competição com os grupos de pesquisa (nunca!) mas actuando como um catalisador.

Começo por tentar delimitar, de forma não erudita, o que é o processamento computacional de uma língua, insistindo em seguida nos problemas identificados para a língua portuguesa. Então, sim, referirei a actividade do presente projecto e as nossas esperanças para o futuro.

2. Processamento computacional de uma língua

O processamento computacional de uma língua pode ser definido como o uso do computador para lidar com a língua (quer «compreendê-a» quer produzi-la). Ora, o conhecimento humano é expresso, quase exclusivamente, em linguagem natural: De facto, é através da língua que lemos e absorvemos informação, e que a transmitimos. É através da língua que ensinamos e aprendemos. Maioritariamente, é através dela que comunicamos com o mundo que nos rodeia. Daí que o processamento de linguagem natural pode ser considerado como a principal forma de manipular o conhecimento humano com o auxílio do computador.

Como disciplina científica, a área do processamento da linguagem natural (também chamada «linguística computacional», «engenharia da linguagem» ou «tecnologia das línguas humanas») tem aplicações múltiplas (veja-se Santos (2001) para uma panorâmica mais abrangente da engenharia da linguagem, precisamente através das suas aplicações). Destacamos aqui, a título de exemplo:

* a procura de informação ("information retrieval") em grandes quantidades de texto – por exemplo na rede(Web)
* a tradução automática e os vários sistemas de apoio à tradução, como memórias de tradução, estações de trabalho dedicadas, etc.
* a ajuda à redacção, incluindo sob esta vasta designação correctores ortográficos, sintácticos, estilísticos, sistemas de crítica à redacção em língua estrangeira através de exemplos, etc.
* a legendagem automática de notícias transmitidas em canais televisivos
* a obtenção (semi)automática de informação sobre domínios especializados
* a criação semi-automática de recursos lexicais, desde dicionários bilingues a tesauros, enciclopédias e redes lexicais
* a concepção de assistentes computacionais que respondam a perguntas e forneçam ajuda não trivial a um utilizador
* os sistemas de ensino (de língua e de outras disciplinas), desde avaliadores semi-automáticos a tutores inteligentes que analisem as causas de erro e a progressão de um determinado aprendente
* a criação automática de resumos e de documentos adaptados a leitores com perfis diferentes
* a obtenção de sistemas personalizados com base na história e nos interesses de cada utilizador
* a indexação inteligente (e correspondente melhoria significativa da procura e da qualidade dos resultados)
* a tradução entre diferentes modalidades (fala - texto) para cidadãos com necessidades especiais ou para trabalhos ou situações que não permitam a utilização da vista ou o recurso a um teclado
* os jogos didácticos – e todo o tipo de entretenimento electrónico em geral – que interagem na língua do utilizador

Não se pense, contudo, que estas aplicações são fáceis de implementar, ou que existem de momento sistemas computacionais que as desempenhem a um nível satisfatório seja para que língua for. Existe, sim, trabalho para atingir estes objectivos, primordialmente para o inglês.

É nossa convicção que um trabalho paralelo terá de existir tendo como alvo o português, visto que não é válida a hipótese ingénua de que "resolvido para uma língua, resolvido para todas". Cada língua é um sistema excepcionalmente complexo com um funcionamento próprio, e é preferível conhecer e saber medir os problemas que se pretende resolver do que adaptar métodos e sistemas de uma língua a outra (desenvolvi este assunto pormenorizadamente em Santos (1999d)).

Antes de prosseguir, darei apenas alguma ideia de que "recursos" se está a falar no presente contexto. Recursos linguísticos não se referem, neste âmbito, a pessoas qualificadas ou línguas "ricas", mas a materiais relacionados com o processamento da língua. Estes podem ser agrupados em alguns grupos:

Corpora - Conjuntos de textos compilados com o fim de responderem a perguntas sobre a língua, ou poderem servir de treino a sistemas que a processam.

Léxicos - Informação sobre palavras e/ou expressões compilada para um fim específico – quer o de documentar a

língua para um utilizador humano (os dicionários) ou documentar o conhecimento num conjunto de áreas (as enciclopédias) ou para ajudar ferramentas a funcionar (tal como listas de palavras para um corrector ortográfico, características sintácticas para dirigir um analisador sintáctico, listas de palavras e suas traduções para ajudar/alimentar a tradução (semi) automática, relações entre palavras ou acepções de forma a guiar um processamento semântico, informação estilística para ser usada na geração de texto, etc. etc.).

Ferramentas - Programas que possam ser reutilizados para o processamento da língua portuguesa, não exigindo que todos comecem de raiz, mas que possam, pelo contrário, aproveitar o trabalho já desenvolvido por outros. Por outras palavras, seria indiscutivelmente benéfico que se pudesse utilizar um analisador sintáctico do português para desenvolver aplicações que necessitem dessa análise sem ter de começar do zero. Ou pelo menos, invocar um analisador morfológico ou um gerador de conjugações verbais sem o programar de raiz.

Recursos para avaliação - Para testar áreas difíceis, sem respostas sim/não, é preciso recorrer a conjuntos de respostas previamente preparadas, compiladas para efeitos de avaliação.

3. Os problemas identificados para a língua portuguesa

Identificada esta área como prioritária, foi lançado em Portugal um debate a nível nacional, por ocasião dos debates sobre política científica que informaram o *Livro Branco para o Desenvolvimento Científico e Tecnológico (1999-2006)* (MCT, Portugal) com base num documento preparatório (Santos, 1999a) para a discussão da área do processamento computacional do português.

Uma das principais hipóteses veiculadas por esse documento era a de que devíamos concentrar esforços no português e não no processamento da língua em geral (ou na linguística computacional). Veja-se Santos (1999b) para uma fundamentação desta escolha.

Outro dos pontos de partida era que, sendo a língua de todos (um bem comum, um direito do cidadão), não se devia deixar – neste campo mais do que em qualquer outro – que alguém (grupo, empresa, academia, etc.) se arvorasse em dono da língua, e que forçasse soluções proprietárias num problema que devia ser de todos – e, portanto, subordinado ao patrocínio e regulação do Estado, enquanto representante do povo.

Por isso, soluções de partilha de recursos, disponibilização dos resultados, uso de ferramentas com código aberto ("open source") e cooperação em vez de competição foram apontados como os modelos a seguir, complementados por um elevado rigor no financiamento e a exigência da avaliação do trabalho alcançado. Assim, na proposta de estabelecimento do centro, ficou bem claro que

> *os recursos criados serão sempre do domínio público, ainda que a sua disponibilização obedeça à preocupação de acautelar os interesses dos eventuais donos ou detentores dos direitos intelectuais, seja através de compensação monetária, seja pelo desenvolvimento de projectos conjuntos de utilidade mútua ou, ainda, recorrendo a outros mecanismos pertinentes.*

Finalmente, era ponto assente que esta área tinha vários problemas graves, enquanto que era certo que o carácter cada vez mais importante que as tecnologias da informação e da comunicação desempenham no nosso dia-a-dia iria obrigar ao uso crescente de sistemas de tratamento automático ou semi-automático da língua, quer falada quer escrita. De facto, só na parte final da década de 90 a tecnologia informática começou a estar num nível de desenvolvimento (capacidade de processamento, miniaturização, custo, ...) que permite uma maior penetração destas tecnologias conducente à sua massificação.

Alguns dos problemas identificados foram: área muito heterogénea, com dois pólos (correspondentes à engenharia e à linguística), com tradições diferentes e sem facilidade de comunicação entre elas, falta de identificação da área como interessante em ambos os campos (é uma área periférica tanto para a informática como para a linguística), falta quase total de pessoas que soubessem de ambos os campos. De facto, equipas interdisciplinares não são suficientes. Como Fernando Pereira muito bem disse na discussão pública, o que é preciso é pessoas interdisciplinares. Além disso, em Portugal havia uma tradição de lutas e guerrilhas entre os vários (e poucos) actores no campo, o que ainda prejudicava mais a possibilidade de cooperação e colaboração para o objectiovo do progresso comum da área. Como será evidente já do que foi escrito, a falta quase total de recursos públicos partilháveis e a falta total de avaliação dos resultados obtidos foram considerados também como problemas sérios.

Claro que muitos destes problemas não admitem uma solução externa; contudo, alguns factores podem ser atenuados e, sobretudo, pode tentar construir-se um futuro diferente, admitindo que os factores estão interligados.

Assim, como mencionado anteriormente, resolvemos concentrar-nos nos capítulos dos recursos e da avaliação, não perdendo contudo de vista a questão da disseminação e informação da área.

Os principais objectivos do centro, neste momento na sua infância – visto que o projecto da sua constituição foi apenas aprovado em 2001 –, podem ser resumidos da seguinte forma:

* facilitar o acesso aos recursos já existentes
* desenvolver de forma harmoniosa e em colaboração com os interessados os recursos considerados mais prementes
* organizar avaliações e encontros científicos que envolvam a comunidade como um todo
* velar por que os recursos encaminhados para esta área possam aproveitar ao máximo o progresso desta
* além de manter e melhorar o portal sobre o processamento computacional do português, http://www.portugues.mct.pt, manter o catálogo de recursos e actores actualizado.

Veremos, na próxima secção, aquilo que já foi feito neste sentido. Mas, mais especificamente, podemos dizer que a actividade do centro reparte-se entre:

1. assegurar dos serviços básicos de repositório, distribuição e catálogo, lançando as bases para tal vir a ser feito de forma distribuída
2. desenvolvimento de alguns recursos pelo próprio centro, sobretudo recursos para avaliação ou para calibragem
3. a organização de "avaliações conjuntas" (conferências de comparação de sistemas onde o método de avaliação é discutido e criado pelos participantes)
4. fornecimento de uma infraestrutura para projectos colaborativos de interesse geral (ou seja, assumir a tarefa da organização quando tal se justificar)

Como resultado indirecto, podemos evidentemente mencionar

5. a formação de pessoal especializado em gestão e disponibilização de recursos

Finalmente, e ainda que inicialmente planeada,

6. a gestão de um programa de desenvolvimento de recursos (incluindo recursos de formação) por concurso público, em presumível articulação com a Fundação para a Ciência e a Tecnologia portuguesa (agência de financiamento) não foi, de facto, implementada, devido à complexidade de todo o processo administrativo exigido.

Note-se que não pensamos que a actividade do centro resolva os problemas da área, levando a um florescimento imediato de aplicações e trabalhos de valor – apenas a consideramos um passo num processo mais vasto, que passa também por um aumento de rigor em toda a área. De facto, o centro de recursos é apenas um dos vectores de uma estratégia mais vasta a nível do Ministério da Ciência e da Tecnologia português na área, que já levou à redefinição do processo de candidaturas dos projectos apresentados à Fundação para a Ciência e a Tecnologia, com a sua avaliação através de discussão pública com um painel de especialistas estrangeiros.

Quanto à falta de transparência, é evidente que tudo fazemos para a contrariar, começando por nós próprios:

> *Além de ser competência do centro facultar informação clara sobre como obter os recursos que o centro mantém ou a que dá acesso, todas as decisões e projectos em que o centro se envolve serão públicos e disponibilizados na rede para consulta de todos os interessados.*

4. Trabalho presente

Conforme descrito em Santos (2000a) e adaptado para o presente artigo, o trabalho do projecto "Processamento computacional do português" e do centro de recursos que lhe deu continuidade pode ser classificado em três vertentes principais:

* disseminação e catalogação da área
* criação e disponibilização de recursos
* avaliação

o que não quer dizer que não existam actividades e tarefas que englobem mais do que uma destas vertentes, como se verá a seguir.

4.1 Disseminação e catalogação da área

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 84 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 (Rocha, 2000) não significa que a sua qualidade esteja assegurada, nem que o único analisador sintáctico mencionado 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 essa razão, além de ter desenvolvido algumas ferramentas internas de gestão do nosso sítio e principalmente do catálogo, implementámos um serviço de procura, o **Busca**, acessível de http://cgi.portugues.mct.pt/Busca/.

O Busca é um sistema de busca sobre o conteúdo do nosso catálogo que permite a um utilizador chegar mais depressa às páginas procuradas. De momento, o Busca resume-se a um sistema de procura cujo "espaço" são todas as páginas apontadas pelo nosso catálogo, com algumas capacidades adicionais de fornecer informação estruturada sobre investigadores relacionados com o processamento da língua portuguesa, dadas as listas de doutorados e de projectos com financiamento público em Portugal a que temos acesso.

É interessante mencionar, contudo, 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 possíveis serão a habituação dos utilizadores à estrutura do catálogo – visto que o sistema de busca foi criado depois de os utilizadores se terem, presumivelmente, adaptado à estrutura do sítio; a sua descrença em relação a ferramentas de procura, muitas vezes pouco cooperativas; ou mesmo a sua preferência pelos grandes motores de procura quando têm uma pergunta específica.

Além disso, é preciso reconhecer que a falta de uma terminologia assente na área também impede que uma procura possa devolver todos os artigos ou páginas relacionados com uma dada questão, dado que praticamente cada pesquisador usa termos diferentes, e que, não sendo conhecidos de antemão, impedem de facto a procura por termos.

Por todas estas razões tencionamos desenvolver este sistema de forma a poder tornar-se realmente útil e cumprir o papel que lhe estava atribuído. Pretendemos torná-lo progressivamente mais inteligente, por um lado detectando automaticamente possíveis candidatos para enriquecer o catálogo e, por outro, evoluindo para um sistema de procura de texto na Web dentro de uma determinada área temática (usando eventualmente categorização semi-automática).

Nessa linha, estamos a iniciar trabalho na procura de informação na Web, tentando obter um modelo da procura real que os utilizadores efectuam de forma a poder estudar formas de melhorá-la usando o processamento de linguagem natural.

Concomitantemente, estamos tentando obter um modelo do utilizador que acede aos nossos corpora através da rede (serviços COMPARA e AC/DC) de forma a melhorar os serviços e compreender melhor qual o uso que lhes é dado, através do estudo dos "logs" (que podemos comparar a pegadas que os utilizadores deixam quando usam um serviço na Web).

Para finalizar a secção referente à disseminação e informação, apresentamos alguns dados concretos quer sobre o nosso sítio na rede quer sobre as visitas recebidas. A tabela 1 e correspondente figura 1 dão uma ideia da evolução da nossa presença na rede em termos do número de páginas que mantemos.

Tabela 1: Número de páginas na rede

Ano	Mês	Páginas
2001	Dezembro	824
	Setembro	642
	Julho	630
	Março	614
2000	Dezembro	599
	Setembro	346
	Julho	328
	Março	321
1999	Dezembro	311
	Setembro	93
	Julho	87
	Março	80
1998	Dezembro	76
	Setembro	16

Tamanho da presença na rede do projeto em número de páginas

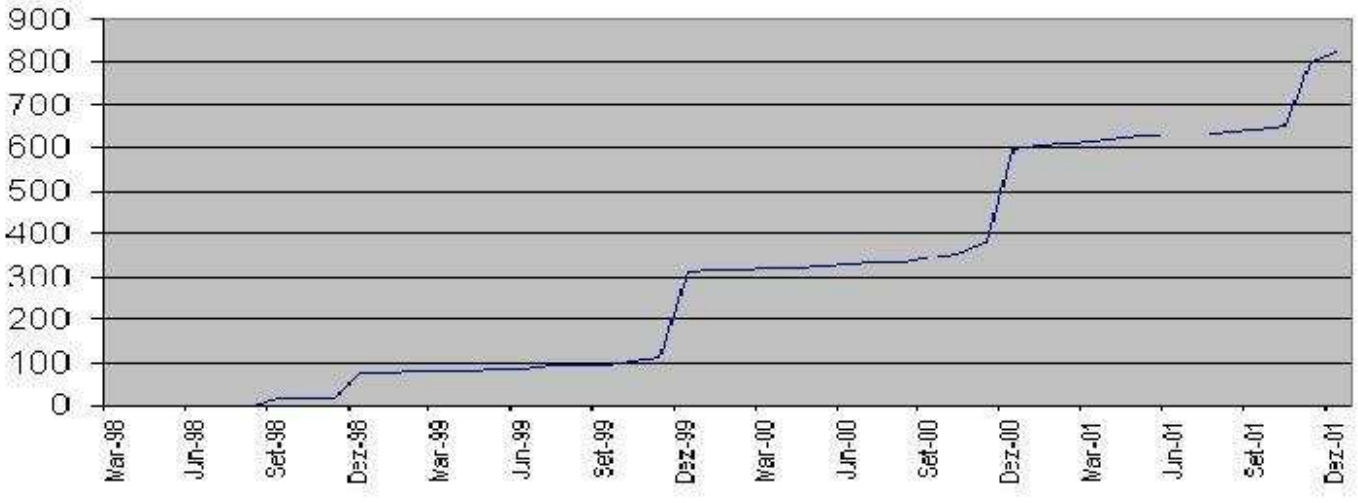
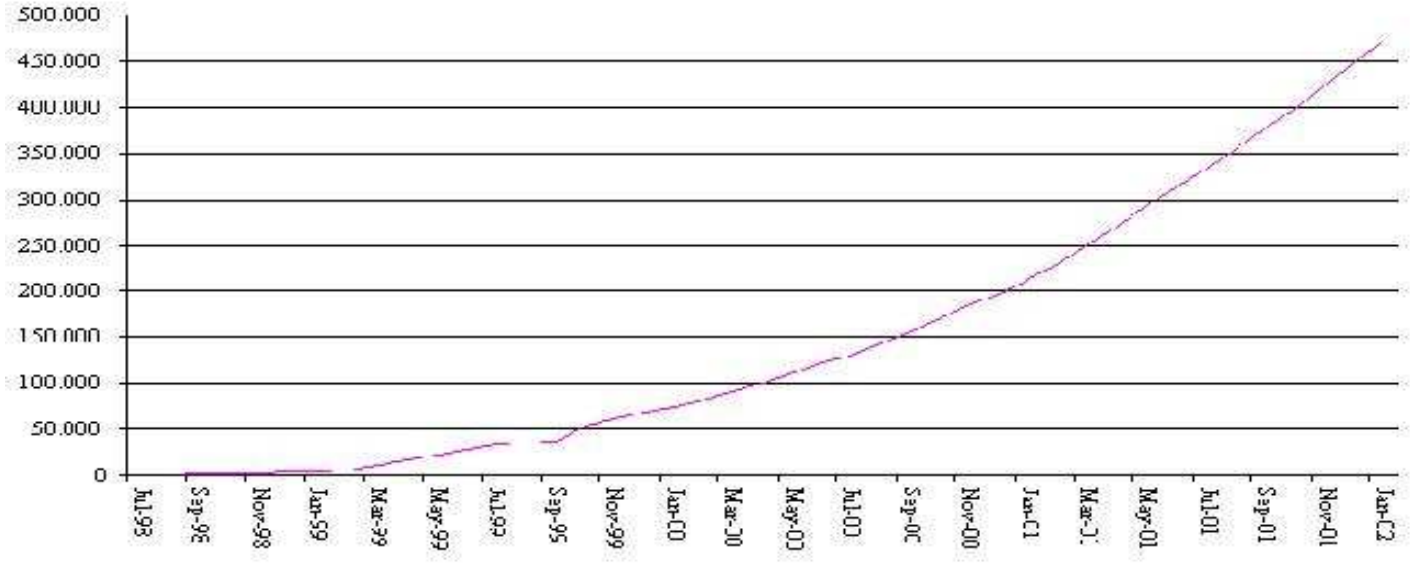


Figura 1: Tamanho do sítio do projecto "processamento computacional do português"

A figura 2, por seu lado, apresenta a evolução do número de visitas ao nosso projecto. A 1 de Fevereiro de 2002, tínhamos tido 471.608 acessos (acumulados) às nossas páginas, provenientes de 89.085 computadores diferentes, 24.667 dos quais tinham interrogado os corpora e apenas 3.400 o sistema de procura.

Número de acessos acumulados às páginas do projeto até 1 de fevereiro de 2002



4.2 Disponibilização de recursos e a sua eventual criação

Por ordem de lançamento temporal pelo projecto "processamento computacional do português", veja-se a série de projectos a que nos temos dedicado no âmbito da disponibilização de recursos:

O projecto AC/DC, acessível de http://cgi.portugues.mct.pt/acesso/, foi lançado na rede a 23 de Setembro de 1999.

Este projecto (cujo nome por extenso é **Acesso a Corpora / Disponibilização de Corpora**) pretende tornar fácil a consulta, para investigação sobre a língua, a grandes quantidades de texto em português, com um mínimo de requisitos técnicos por parte do utilizador. Esta linha de acção foi sugerida em Santos (1999c) como forma de contrariar a grande dificuldade, então identificada, de acesso a este tipo de material. O projecto AC/DC faculta o acesso através da Web a diversos corpora, codificados e processados de forma a serem enriquecidos com informação pertinente para estudos de vária índole: Assim, esses textos contêm associada indicação sobre a segmentação (separação em frases, parágrafos e outras unidades estruturais) e o tipo de texto (jornalístico, ensaio, texto literário, didáctico, etc.), assim como, para cada palavra, a categoria morfossintáctica e informação morfológica, e a função sintáctica (num formalismo de gramática dependencial), etc. (Veja-se Santos & Bick (2000) para uma descrição técnica aprofundada.)

Tal permite a um linguista (ou a um engenheiro da linguagem) interrogar um corpus formulando questões que excedem largamente a simples procura de palavras ou formas, bastando para isso ter acesso à Internet.

De momento, já se encontram interrogáveis textos perfazendo mais de 230 milhões de palavras (de português de Portugal e do Brasil), incluindo texto jornalístico (a grande maioria), literário, técnico, didáctico, assim como ensaio e correspondência. O trabalho prossegue com a integração de mais material textual, em curso.

No momento presente, temos o material acessível para consulta descrito na tabela 2:

Tabela 2: Breve descrição dos corpora acessíveis através do projecto AC/DC

Corpus	Tamanho (unidades)	Tamanho (palavras)	Tamanho (frases)	Breve descrição
natura natpanot	7.257.408 7.325.573	6.256.880 6.293.686	226.118 225.969	Texto jornalístico, PÚBLICO, Portugal, 1991-1994, dois parágrafos por edição

enpcpub enpcanot	89.865 90.508	72.244 72.370	4.369 4.369	Literatura traduzida do inglês, de 5 obras do corpus ENPC
minho minhanot	2.084.088 1.928.976	1.738.127 1.596.316	53.233 47.588	Artigos de jornal regional, Diário do Minho, Portugal, antes da revisão
eci-ebr ebranot	891.835 898.281	721.951 722.396	45.617 44.627	Texto brasileiro, do corpus Borba-Ramsey, compilado pelo ECI
eci-ee eeanot	30.157 30.968	26.515 26.993	780 772	Texto de chamada do programa europeu ESPRIT, em português de Portugal
saocarlos scanot	34.601.701 32.812.673	27.244.237 25.739.598	1.372.396 1.271.377	Texto do corpus NILC, em português brasileiro, contendo maioritariamente texto jornalístico, mas também cartas comerciais e textos didácticos
frasespp fppanot	19.340 19.548	16.225 16.203	594 594	Frases em português de Portugal
frasespb fpbanot	22.488 22.733	19.155 19.164	652 652	Frases em português do Brasil
cprmi cpprmianot	1.197.908 1.198.733	997.600 995.955	38.348 38.258	Texto jornalístico, PÚBLICO, Portugal, 1991-1998, extractos de dois parágrafos
ancib ancibanot	357.545 363.360	287.588 291.678	12.125 12.068	Correio electrónico em português brasileiro correspondente ao tráfego na lista da ANCIB
cetempublico cetempanot	229.038.019	191.687.833	7.082.094	Texto jornalístico, PÚBLICO, Portugal, 1991-1998, extractos de dois parágrafos

O projecto **COMPARA/DISPARA**, acessível de http://www.portugues.mct.pt/COMPARA/, foi lançado em Julho de 2000.

O projecto COMPARA/DISPARA é um congénere do AC/DC para textos paralelos em português e inglês. O corpus COMPARA foi, contudo, criado de raiz (a escolha dos textos e a obtenção da permissão de inclusão no corpus é da responsabilidade exclusiva de Ana Frankenberg-Garcia). A problemática de interrogar um texto e a sua tradução levou ao desenvolvimento de uma interface especial para corpora paralelos, o DISPARA. Este projecto encontra-se em progresso: temos já autorização para incluir cerca de uma centena de textos, facultando de momento treze pares de textos para consulta através da rede. (Veja-se Frankenberg-Garcia & Santos (no prelo a; no prelo b) para mais informação.)

O projecto do **CETEMPúblico**, veja-se http://cgi.portugues.mct.pt/cetempublico/, levou à disponibilização da primeira versão deste corpus a 25 de Julho de 2000.

O CETEos: além de dar acesso através da rede (este corpus também está, evidentemente, acessível através do projecto AC/DC), quisemos poder distribuir fisicamente o texto todo. Afinal de contas, o público alvo da nossa acção são precisamente investigadores e engenheiros que desenvolvem ferramentMPúblico (Corpus de Extractos de Textos Electrónicos MCT/Público) foi um recurso criado com o objectivo de ir mais além na facilitação de recursas que lidam com o português – e que precisam, portanto, além de consultar, de manipular e utilizar corpora nos seus programs. O texto, do jornal diário português PÚBLICO, foi dividido em extractos ordenados aleatoriamente, de forma a impedir a reconstrução das notícias integrais, e é distribuído em CD (versão 1.0) gratuitamente a todos quantos registarem o seu endereço postal, além de ser distribuído (na sua versão 1.7) pelo Linguistic Data Consortium, a maior agência de distribuição de recursos linguísticos. À data de redacção do presente texto (Fevereiro de 2002), já tivemos mais de 200 pedidos directos oriundos de pessoas ou grupos espalhados pelo mundo inteiro. (Veja-se Rocha & Santos (2000) para a descrição da sua criação, e Santos & Rocha (2001) para uma primeira avaliação deste recurso.)

Um projecto paralelo focando a linguagem jornalística brasileira, desenvolvido em colaboração com o Núcleo Interinstitucional de Lingüística Computacional (NILC) da Universidade de São Paulo, do Brasil, encontra-se em fase final de execução.

Finalmente, o projecto da **Floresta Sintá(c)tica**, veja-se http://cgi.portugues.mct.pt/treebank/, tem como objectivo criar um conjunto de unidades correctamente analisadas sintacticamente (revistas por linguistas) que possam servir tanto como instrumento de calibragem de analisadores sintácticos automáticos, como para melhorar a cobertura descritiva da análise sintáctica computacional, obrigando a um consenso ou pelo menos a uma formalização das escolhas e alternativas tomadas em áreas "cinzentas" da sintaxe.

Se é, por um lado, possível considerá-lo como a continuação natural do projecto AC/DC (por fornecer uma informação mais fina – árvores sintácticas – e com um nível de correcção consideravelmente superior – devido à revisão humana), o projecto Floresta Sintá(c)tica é, por outro lado, mais ambicioso, tanto em termos de concepção científica como em termos políticos. Ou seja, não só o processo de criação do banco de árvores é original, mas também é nossa intenção congregar, após uma primeira fase de experimentação ligada a um analisador sintáctico determinado (o PALAVRAS de Eckhard Bick, veja-se Bick, 2000), fase essa descrita em Afonso et al. (2002), todos os grupos que se dedicam ao processamento da nossa língua e que aceitem colocar o formato do seu analisador para escrutínio do resto da comunidade.

No primeiro ano da sua actividade, correspondendo à primeira fase, foram criadas 1.427 árvores (totalizando aproximadamente 35.000 palavras), que podem ser obtidas em ficheiro texto, além de interrogadas através da rede usando o sistema Águia desenvolvido pelo nosso projecto, além de ter sido investido um esforço considerável na documentação.

De momento, estamos a considerar qual a melhor forma de um seu prolongamento, após uma reflexão sobre a utilidade, a qualidade do produto obtido, e a possibilidade de vir a incorporar elementos de grupos diferentes e que sigam paradigmas linguísticos diferentes.

4.3 Avaliação

A avaliação exige um conhecimento elevado do problema e o desenvolvimento de metodologias próprias. É, pois, uma tarefa vastíssima, mas aquela que nos parece requerer a maior atenção do ponto de vista do estudo e desenvolvimento da nossa área. Cito, a esse propósito, Hirschmann (1998b:302, tradução minha), que afirma:

> *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.*

Como abordado no tutorial sobre avaliação que preparei há dois anos (Santos, 2000b), existem variadíssimas formas de avaliar, mesmo quando o objectivo é uma simples avaliação técnica e não na óptica do utilizador, do financiador ou do mercado. Desde a comparação de vários produtos segundo um conjunto de parâmetros externos (como por exemplo Rocha (2000)) a uma análise da contribuição de alguns factores no desempenho de uma única ferramenta (Santos & Oksefjell, 2000), até à análise de recursos (Santos & Rocha, 2001) e da repetibilidade da avaliação de hipóteses linguísticas (Santos & Oksefjell, 1999), muitas destas análises podem ser feitas como um trabalho isolado.

Não é, contudo, nesse tipo de avaliação que nós estamos principalmente interessados, mas sim num que leve a um consenso e uma maior participação de toda a comunidade. Estamos, sim, empenhados em organizar uma primeira avaliação conjunta para o português, com pontapé de saída no PorTAL (Junho de 2002) e realização efectiva no PROPOR (Maio de 2003).

O que é uma avaliação conjunta? A minha tradução do termo inglês "evaluation contest" tenta pôr a ênfase numa actividade conjunta e não na competição, mas basicamente refere-se a um acontecimento científico-tecnológico cujo fim é a avaliação dos sistemas que professam executar uma dada tarefa, que pode ser tão simples como a análise morfológica (veja-se a Morpholympics, Hausser, 1994) ou tão complexa como a sumarização (Mochizuki & Okumura, 2000), passando pela análise sintáctica (Black et al., 1991), pela resposta automática a perguntas escritas (Voorhes & Tice, 2000) ou pela interacção falada com sistemas automáticos de venda de passagens de avião (Hirschmann, 1998b).

A forma como se processa este tipo de encontro é a seguinte:

* os participantes reúnem-se e definem um método de avaliação (de preferência baseado em recursos públicos ou partilhados, que possam ser inspeccionados pelo público em geral e não só pelos próprios participantes)
* esse método de avaliação define também uma calendarização e a forma de organizar a conferência sem viciar o processo ou favorecer alguns candidatos

* é dado a todos os sistemas participantes o mesmo conjunto de problemas a resolver, no mesmo prazo
* os resultados são comparados por um grupo de juízes previamente treinados, e os resultados são tornados publicamente acessíveis (ainda que não necessariamente discriminados por quem é quem, se tal tiver sido combinado inicialmente).

É sabido que, por mais que se tentem prever todas as questões com antecedência, todos os processos de avaliação são suficientemente complexos para necessitarem de rodagem e de várias conferências em anos sucessivos até se conseguir um elevado grau de objectividade. Mas, como em todas as questões complexas, sem fazer não há progresso.

5. Esperança no futuro

Gostaríamos que, daqui a vinte anos, o presente projecto/centro pudesse ser identificado, pela comunidade da lusofonia computacional, como tendo propiciado uma viragem significativa na área, tendo acelerado o seu progresso.

Para isso, é preciso que os recursos sejam de facto utilizados na prática, que as equipas colaborem na definição de métodos de avaliação em conjunto, na identificação de problemas resolvidos e não resolvidos, e na criação de recursos para avaliação.

É também preciso que cada vez mais pesquisadores se interessem pela área, querendo dedicar-lhe o seu tempo, não como a um passatempo, mas como actividade principal, e que os decisores, políticos e empresariais, queiram investir nesta disciplina como forma de melhorar o quotidiano de todos os falantes da língua portuguesa.

Agradecimentos

Agradeço a todos membros do projecto, presentes e passados, a discussão em torno da actividade e, claro, o trabalho realizado. Por ordem alfabética, Rachel Aires, Tom Funcke, Renato Haber, Pedro Moura, Signe Oksefjell, Paulo Alexandre Rocha e Alexsandro Santos Soares.

REFERÊNCIAS BIBLIOGRÁFICAS

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. No prelo. "Floresta sintá(c)tica: um treebank para o português". *Actas do XVII Encontro da Associação Portuguesa de Linguística* (Lisboa, Outubro de 2001), APL, 2002, http://www.portugues.mct.pt/Diana/download/AfonsoetalAPL2002.rtf.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press, 2000.

Black, E., S. Abney, D. Flickinger, C. Gdaniek, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini & T. Strzalkowski. 1991. "A procedure for quantitatively comparing the syntactic coverage of English grammars", *Proceedings of the February 1991 DARPA Speech and Natural Language Workshop* (Pacific Grove, CA, February 1991), pp. 306-311.

Frankenberg-Garcia, Ana & Diana Santos. No prelo a. "Introducing COMPARA: the Portuguese-English Parallel Corpus". In *Proceedings of The Second International Conference on Corpus Use and Learning to Translate (CULT2000)* (Scuola Superiore di Lingue Moderne per Interpreti e Traduttori da Università di Bologna, Bertinoro, Italy, November 3-5, 2000), St. Jerome, no prelo, http://www.portugues.mct.pt/Diana/download/FGSantosCULT.rtf.

Frankenberg-Garcia, Ana & Diana Santos. No prelo b. "Apresentando o COMPARA, um corpus português-inglês na Web", Cadernos de Tradução, Universidade de São Paulo, http://www.portugues.mct.pt/Diana/download/FGSantosCadTrad.rtf.

Hausser, Roland. 1994. "The coordinator's final report on the first Morpholympics". *LDV-Forum* **11**(1), 1994, pp. 54-64.

Hirschman, Lynette. 1998a. "Language Understanding Evaluations: Lessons Learned from MUC and ATIS". 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. 1, pp.117-122.

Hirschman, Lynette. 1998b. "The evolution of Evaluation: Lessons from the Message Understanding Conferences".

Computer Speech and Language **12** (4), pp.249-62.

Livro Branco do Desenvolvimento Científico e Tecnológico Português (1999-2006). 1999. Observatório das Ciências e das Tecnologias, Ministério da Ciência e da Tecnologia, http://www.mct.pt/Livro-BrancoCT/.

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

Mochizuki, Hajime & Manabu Okumura. 2000. "A Comparison of Summarization Methods Based on Task-based Evaluation". In Maria Gavriladou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhaouer (eds.), *Proc. Second International Conference on Language Resources and Evaluation*, LREC'2000 (Athens, 31 May - 2 June 2000), Vol 2, pp. 633-39.

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, http://www.portugues.mct.pt/Diana/download/recursos.ps.

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 & Diana Santos. 2000. "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa". In Maria das Graças Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR'2000* (Atibaia, São Paulo, 19 a 22 de Novembro de 2000), pp. 131-140, http://www.portugues.mct.pt/Diana/download/RochaSantosPROPOR2000.rtf.

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

Santos, Diana. 1999b. "Porquê processamento computacional do português e não processamento de linguagem natural?", 24 de Março de 1999, http://www.portugues.mct.pt/branco/Porque.html.

Santos, Diana. 1999c. "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, http://www.portugues.mct.pt/Diana/download/WLC.rtf.

Santos, Diana. 1999d. "Toward Language-specific Applications". *Machine Translation* **14** (2), June 1999, pp.83-112. Kluwer Academic Publishers.

Santos, Diana. 2000a "O projecto Processamento Computacional do Português: Balanço e perspectivas". In Maria das Graças Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR'2000* (Atibaia, São Paulo, 19 a 22 de Novembro de 2000), pp. 105-113, http://www.portugues.mct.pt/Diana/download/SantosPROPOR2000.rtf.

Santos, Diana. 2000b. Tutorial sobre "Evaluation of Natural Language Processing systems". *Joint International Conference IBERAMIA/SBIA 2000* (Atibaia, São Paulo), 19 de Novembro 2000, http://www.portugues.mct.pt/Diana/download/TutEval2000.ps.

Santos, Diana. 2001. "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, pp. 229-259, http://www.portugues.mct.pt/Diana/download/aplicacoes.rtf.

Santos, Diana & Eckhard Bick. 2000. "Providing Internet access to Portuguese corpora: the AC/DC project". In Maria Gavriladou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhaouer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000* (Athens, 31 May - 2 June 2000), pp. 205-210, http://www.portugues.mct.pt/Diana/download/SantosBickLREC2000.rtf.

Santos, Diana & Signe Oksefjell. 1999. "Using a Parallel Corpus to Validate Independent Claims". *Languages in Contrast* Vol. **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, http://www.portugues.mct.pt/Diana/download/SantosOksefjellNodalida99.rtf.

Santos, Diana & Paulo Rocha. 2001. "Evaluating CETEMPúblico, a free resource for Portuguese", *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (Toulouse, 9-11 July 2001), pp.442-449, http://www.portugues.mct.pt/Diana/download/SantosRochaACL2001.rtf.

Veiga, Pedro & Diana Santos. 2001. "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 (Actas do colóquio de 25 a 26 de Janeiro de 2001)*, Lisboa: Edições Colibri, pp. 103-109, http://www.portugues.mct.pt/Diana/download/VeigaSantos.rtf.

Voorhes, Ellen M. & Dawn M. Tice. 2000. "The TREC-8 Question Answering Track". In Maria Gavriladou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhaouer (eds.), *Proc. Second International Conference on Language Resources and Evaluation, LREC 2000* (Athens, 31 May 2000), Vol III, pp.1501-8.

Sobre a autora / About the Author:

Diana Santos
Diana.Santos@sintef.no
Doutora em Engenharia Informática
Investigadora no SINTEF Telecommunications and Informatics, Oslo
Pb 124, Blindern
NO-0314 Oslo, Noruega
http://www.portugues.mct.pt/Diana/