Linguateca:

Relatório relativo ao período 2000-2003

Diana Santos

Relatório referente a:

Projecto POSI - Eixo Prioritário 1: Desenvolver Competências / Medida 1.3 Investigação e Desenvolvimento
Projecto: Centro de Recursos Distribuído para o Processamento Computacional da Língua Portuguesa
Nº de Origem: 4/1.3/C/NRE
Período: 2000-2003

Nota:

Aproveitámos o ensejo para apresentar a actividade da Linguateca no contexto da sua acção continuada, dado que este projecto evoluiu directamente do anterior projecto Processamento Computacional do Português, e continua ao abrigo de outros projectos POSI. Não faria, portanto, sentido segmentar a actividade e apresentá-la apenas até 15 de Maio de 2003. O presente relatório abrange assim toda a actividade atë à data da sua escrita (Setembro de 2003) e apresenta também, embora de passagem, o trabalho começado antes do estabelecimento formal do projecto.

Agradecimento:

O trabalho aqui descrito corresponde a um trabalho de equipa, e como tal é apresentado. Na redacção do presente relatório, contudo, é preciso agradecer a colaboração do Luís Costa e do Luís Sarmento, respectivamente responsáveis pela medição das estatísticas relativas ao sítio da Linguateca, e pela criação das figuras.

Apresentação

O presente texto tem como objectivo principal descrever a actividade da *Linguateca* no período coberto pelo projecto acima, ou seja, de 15 de Maio de 2000 a 14 de Maio de 2003, o qual foi ainda lançado com o nome provisório de *Centro de Recursos – distribuído – para a Língua Portuguesa* (CRdLP).

O que é a Linguateca?

A Linguateca consistiu na materialização de algumas das propostas feitas pelo anterior projecto *Processamento Computacional do Português*[1] (15 de Maio de 1998 a 14 de Maio de 2000), após a identificação dos problemas maiores que assolavam a área e uma primeira demonstração concreta da forma de os resolver.

A Linguateca foi, pois, criada como uma estrutura de disseminação e agilização da construção e disponibilização de recursos para o processamento (computacional) da língua portuguesa, dedicada a três objectivos fundamentais:

- divulgação e catalogação do processamento computacional do português na rede
- disponibilização, melhoria e criação de recursos
- avaliação da área, primordialmente através da organização de avaliações conjuntas

que denominamos como modelo IRA - **I**nformação, **R**ecursos e **A**valiação.

Após estes três anos de actividade, a Linguateca tem a arquitectura representada na figura 1. Do ponto de vista organisatório, a Linguateca foi concebida como uma organização virtual (distribuída) em que os seus pólos criam novos serviços ou recursos para a comunidade, integrando e potenciando o espírito e a filosofia do projecto inicial em centros de investigação de renome e com experiência em processamento do português, através de uma colaboração prática, desenvolvendo serviços e disseminando os recursos já existentes.

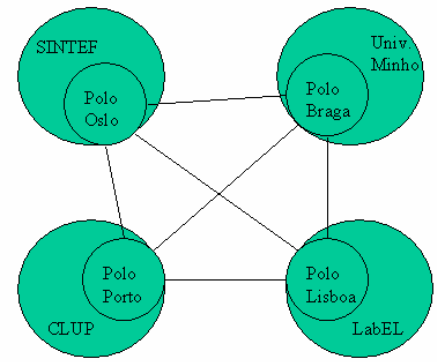


Figura 1: Arquitectura da Linguateca: Quatro pólos integrados em realidades diferentes, com um mesmo objectivo comum

Tal implicou, nestes três anos, por um lado, a actividade continuada de gestão e desenvolvimento dos serviços inicialmente desenvolvidos pelo projecto Processamento Computacional do Português em Oslo (a partir de agora referido simplesmente como pólo de Oslo), e, por outro, a criação de novos serviços e realidades de apoio e de dinamização através do lançamento da actividade dos novos pólos.

[1] contratado pela Agência de Inovação portuguesa ao SINTEF Oslo

Pressupostos básicos da Linguateca

A criação da Linguateca teve como pressupostos as seguintes convicções:

- É preciso uma estrutura que ajude a documentar, criar e armazenar o trabalho feito de forma dispersa na área. Os investigadores não têm condições, quer financeiras, quer de tempo, para distribuir, manter e documentar os recursos que possam vir a criar no âmbito dos projectos em que participam.
- As agências de distribuição multinacionais não têm ocasião, nem possibilidades de controlo de qualidade, comparáveis às de um organismo dedicado ao português e gerido por falantes nativos da nossa língua.
- Recursos públicos são essenciais para o progresso na área, tanto para comparação de métodos diferentes, como para evitar que os investigadores repitam os esforços de colecção de dados e tratamento de direitos que são extremamente morosos.
- Em vez de apostar em organizações regionais, devemos juntar todos os que trabalham com a língua portuguesa, em todas as suas variantes, e daí tentar fomentar uma colaboração estreita entre Portugal, Brasil e os outros países lusófonos, assim como trabalhar com todos os investigadores que se dediquem ao português, independentemente da sua localização geográfica.
- Não há razão para tratar diferentemente empresas e universidades. O trabalho de uns e de outros é igualmente necessário para a melhoria dos produtos e do conhecimento sobre o processamento computacional da nossa língua.

Público-alvo

Podemos identificar três níveis de audiência no processo de melhorar significativamente a situação da população informatizada falante de português, que, cada dia, abrange um número maior de pessoas:

1. a comunidade científica na área do processamento computacional da língua (área também denominada linguística computacional, engenharia da linguagem ou processamento de linguagem natural)
2. os desenvolvedores de programas, serviços ou produtos que integram ou se apoiam em componentes significativos de processamento de português
3. os utilizadores desses programas, que poderão vir a abranger a maioria da comunidade lusofalante

É importante deixar bem claro que o público-alvo da Linguateca se encontra nos dois primeiros níveis, embora o seu resultado terminal se espera que seja o de, num futuro relativamente próximo, melhorar a situação da população em geral no que se refere ao acesso e produção da informação em português.

Principais actividades

Nestes três anos de actividade, foram estabelecidos três novos pólos (a partir do pólo seminal de Oslo) e uma identidade foi criada em torno do nome *Linguateca* (que substituiu, com vantagem, a denominação pouco mnemónica de *CRdLP*):

- O Pólo de Braga (criado em conjunto com o projecto Natura[2] do Departamento de Informática da Universidade do Minho, liderado por José João Dias de Almeida e Pedro Rangel Henriques) foi estabelecido em Novembro de 2000.
- O Pólo de Lisboa (criado em conjunto com o LabEL, Laboratório de Engenharia da Linguagem[3], do Instituto Superior Técnico, liderado por Elisabete Ranchhod) foi estabelecido em Setembro de 2001.
- Pólo do Porto (criado em conjunto com o CLUP, Centro de Linguística da Universidade do Poro, da FLUP, Faculdade de Letras da Universidade do Porto, vertente de terminologia e tradução, liderada por Belinda Maia) foi estabelecido em Outubro de 2002.

Muito resumidamente, a actividade da Linguateca nestes três primeiros anos foi norteada pelos três eixos de actividade complementares já mencionados, a saber: **Disseminação e catalogação** (com a manutenção de um portal na rede constantemente actualizado); **Criação de recursos** (com a sua consequente manutenção e apoio ao utilizador) e **Avaliação** (com a criação de um sítio dedicado à avaliação conjunta e a organização da primeira avaliação conjunta para o português).

Estrutura do relatório

No presente relatório pretendemos fornecer algumas medidas objectivas do resultado da nossa actuação, nestes três eixos. O texto encontra-se, portanto, dividido em três secções principais, dedicadas cada uma à sua vertente, finalizando com uma apresentação da equipa e das publicações e outra documentação produzidas no âmbito da Linguateca.

[2] Veja-se o endereço do projecto Natura: http://natura.di.uminho.pt.
[3] Veja-se o endereço do LabEL: http://label.ist.utl.pt.

Eixo 1: Disseminação e informação

Nestes três anos, desenvolvemos e melhorámos o portal sobre o processamento computacional do português, que atingiu no princípio de Agosto de 2003 um milhão de visitas, e que tem as seguintes categorias principais:

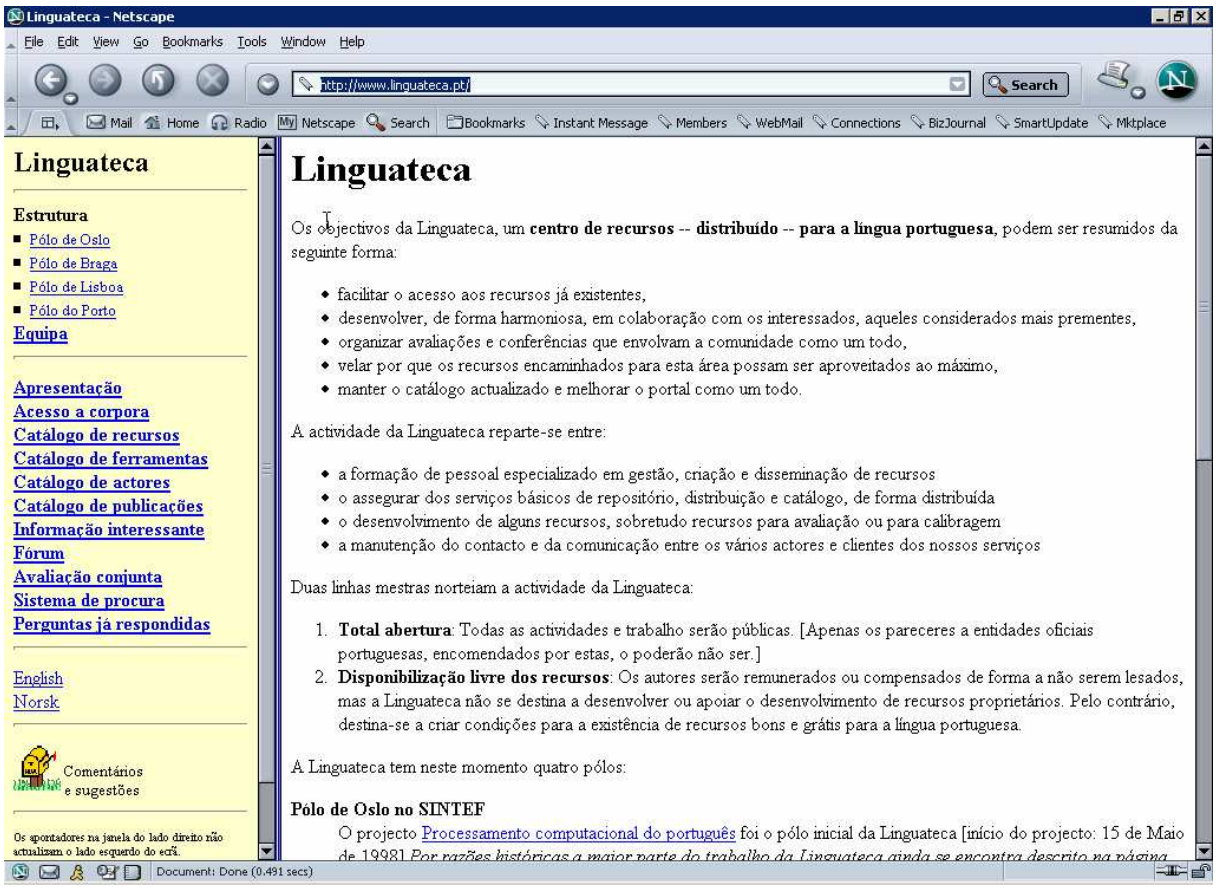


Figura 2: A página de entrada da Linguateca

Catálogo de recursos: informação sobre recursos linguístico-computacionais para o português: corpora, léxicos, textos, serviços – 743 entradas

Catálogo de ferramentas: informação sobre programas, ferramentas e ambientes de trabalho para o processamento computacional do português – 128 entradas

Catálogo de actores: informação sobre projectos, grupos, e investigadores – 395 entradas

Catálogo de publicações: bibliografia constantemente actualizada sobre o que se escreve e publica na área – 750 entradas

Fórum: informação sobre conferências, notícias, cursos e bolsa de emprego na área do processamento computacional do português, junto com um arquivo das entradas anteriores (até agora noticiámos 295 peças)

Informação interessante: conjunto de ligações relevantes ou associadas ao processamento do português, tais como informação jurídica, informação de gestão da área, informação técnica, etc. – 135 entradas

Avaliação conjunta: informação e portal dedicado à organização de avaliações conjuntas para o português, que será descrito na parte 3 do presente texto.

Além disso, o portal da Linguateca dá acesso directo aos recursos criados no seu âmbito, assim como a informação do próprio projecto, como estrutura e equipa da Linguateca, informação sobre as nossas visitam, resposta a perguntas pertinentes sobre a nossa actividade, lista de publicações e informação técnica.

Algumas medidas do impacto e da grandeza do portal

Apresentamos algumas estatísticas sobre a nossa presença na rede que permitem avaliar objectivamente o impacto da nossa actividade na vertente do portal.

Quanto ao tamanho do portal, apresentamos na tabela 1 a quantidade de material envolvido, por tipo de recurso.

Tabela 1: Distribuição do conteúdo do sítio, por tipo de ficheiro

Tipo	**Quantidade**
html	1043
ps	110
txt	120
rtf	52
doc	45
pdf	8
Total	**1378**

O número de ligações para fora (URLs externas para que apontamos, contando com os catálogos e também com outras referências) atinge os 2439, como indicado na tabela 2.

Tabela 2: Ligações externas, por localização geográfica

Sufixo geográfico	Quantidade
`pt`	804
`br`	705
`com`	284
`org`	148
`edu`	111
`de`	59
`uk`	51
`no`	35
`fr`	33
`net`	29
`dk`	26
outros	154
Total	**2439**

Por outro lado, é revelador do intenso esforço de informação o facto de que, das 1043 páginas HTML que servimos ao público, um total de 782 correspondem a documentação própria, sem contar com as páginas dos diversos catálogos, que são evidentemente redigidas pela equipa da Linguateca também.

Podemos com estes valores indicar que o nosso sítio na rede pode ser considerado de grande dimensão em comparação com a Web portuguesa.[4]

Daí que não admira termos atingido recentemente um milhão de visitas[5]. Sobre aquelas que indicam a proveniência (e que correspondem a cerca de 50% do total), calculámos a repartição geográfica, que apresentamos na figura seguinte. Note-se que, embora os países de língua portuguesa sejam, evidentemente, os responsáveis pela maior parte das visitas, temos um número interessante de visitantes a nível internacional, da ordem de um quinto de todas as visitas com assinatura reconhecida.

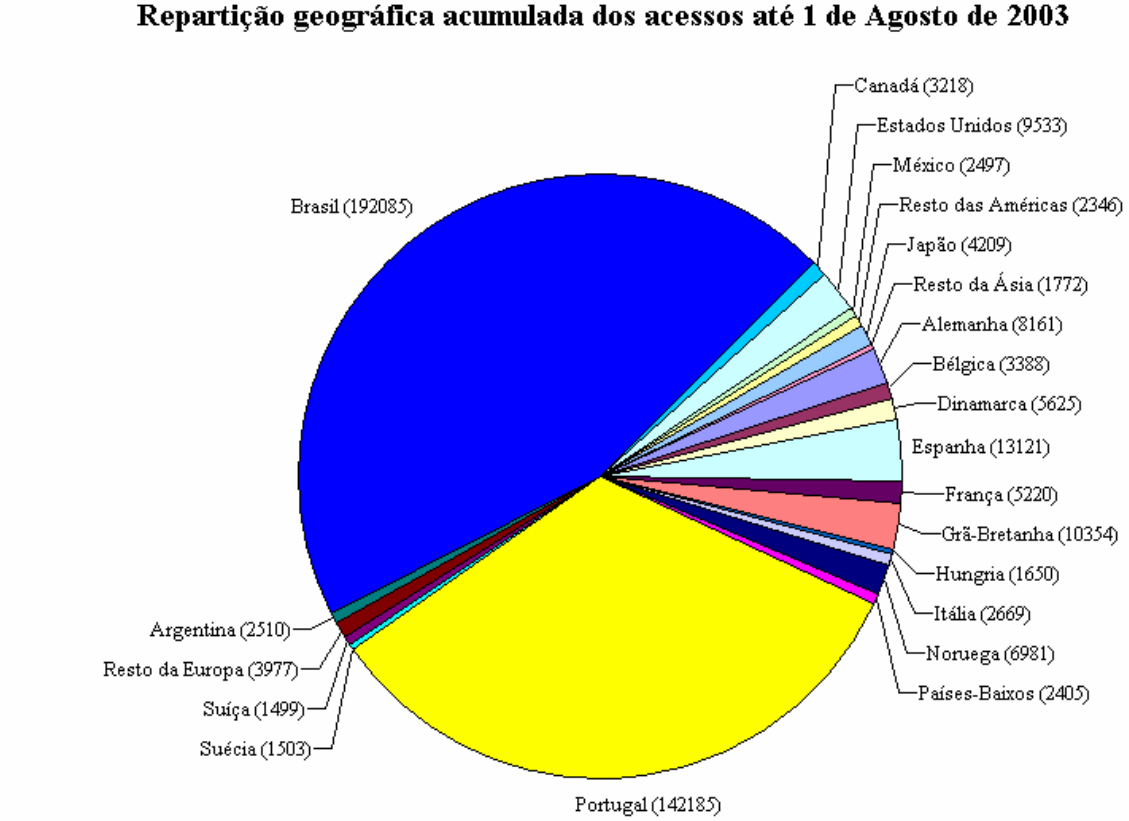


Figura 3: A distribuição geográfica dos acessos que permitem conhecer a origem (cerca de 50% do total)

Outra forma, complementar, de avaliar o impacto da nossa actividade de disseminação é medir o número de ligações na rede para dentro do nosso sítio (o que representa o núemro de páginas sítios que apontam para nós, de que temos conhecimento), e de que encontrámos, à data da escrita do presente relatório, serem da ordem das 600.

Outro indicador é o número médio de visitas mensais, na ordem das 30 mil, e de que apresentamos o gráfico desde o início da nossa presença na rede, na figura 4.

[4] Para esta apreciação, usámos o trabalho de Daniel Gomes e Mário J. Silva, "A Characterization of the Portuguese Web", *3rd ECDL Workshop on Web Archives*, Trondheim, Noruega, Agosto 2003.

[5] Por razões técnicas associadas à configuração dos servidores em Lisboa e no Porto, os números que apresentamos são apenas de Oslo e Braga.

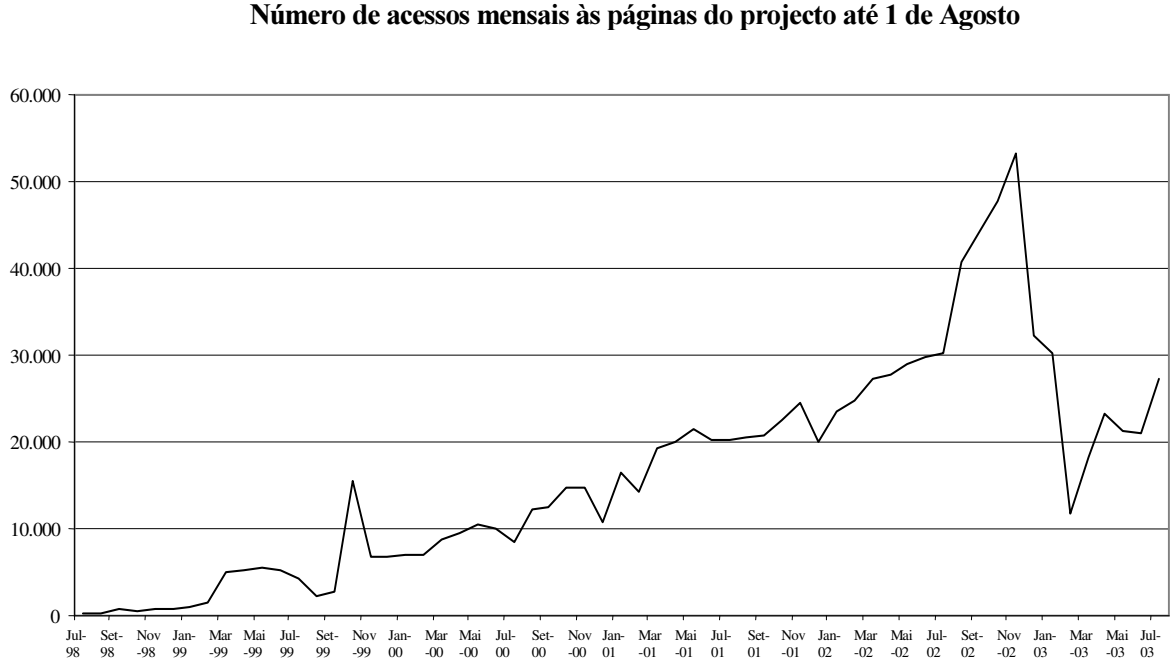


Figura 4. O número de visitas mensal ao sítio da Linguateca (antes projecto Processamento Computacional do Português) desde a sua criação

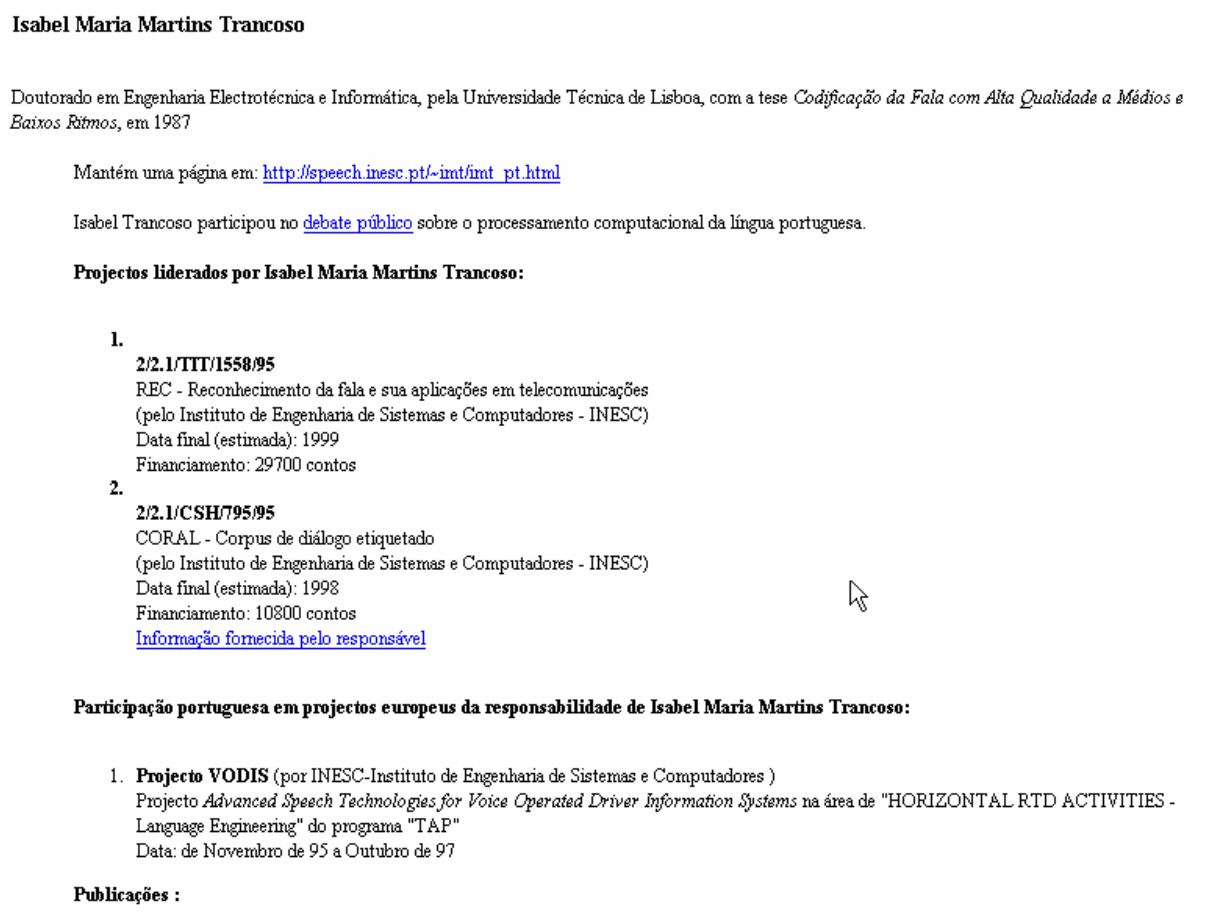
Outros indicadores, relativos aos serviços na rede, aos recursos encomendados através de nós, à nossa actividade de documentação e à participação nas actividades de avaliação conjunta, serão apresentados nas respectivas secções.

Busca: Sistema de procura no portal

A Linguateca desenvolveu um sistema de busca no seu portal, com duas formas de procura: procura geral, na forma dos motores de busca na rede, e busca por pessoas. Nesta última, alguma atenção foi dada às peculiaridades da língua portuguesa e em particular dos seus nomes, assim como à apresentação da informação, com vários patamares (apenas apresentados se tivermos alguma informação nessa rubrica):

- apresentação de página pessoal (se a tivermos no catálogo)
- apresentação de participação no debate sobre o futuro da área (se participou)
- apresentação de liderança de projectos
- indicação do título da tese (se presente nas nossas listas)
- apresentação das publicações
- procura geral do nome em todas as páginas

A figura 5 apresenta o resultado de uma procura por pessoa.



Isabel Maria Martins Trancoso

Doutorado em Engenharia Electrotécnica e Informática, pela Universidade Técnica de Lisboa, com a tese *Codificação da Fala com Alta Qualidade a Médios e Baixos Ritmos*, em 1987

Mantém uma página em: http://speech.inesc.pt/~imt/imt_pt.html

Isabel Trancoso participou no debate público sobre o processamento computacional da língua portuguesa.

Projectos liderados por Isabel Maria Martins Trancoso:

1. **2/2.1/TIT/1558/95**
   REC - Reconhecimento da fala e sua aplicações em telecomunicações
   (pelo Instituto de Engenharia de Sistemas e Computadores - INESC)
   Data final (estimada): 1999
   Financiamento: 29700 contos
2. **2/2.1/CSH/795/95**
   CORAL - Corpus de diálogo etiquetado
   (pelo Instituto de Engenharia de Sistemas e Computadores - INESC)
   Data final (estimada): 1998
   Financiamento: 10800 contos
   Informação fornecida pelo responsável

Participação portuguesa em projectos europeus da responsabilidade de Isabel Maria Martins Trancoso:

1. **Projecto VODIS** (por INESC-Instituto de Engenharia de Sistemas e Computadores)
   Projecto *Advanced Speech Technologies for Voice Operated Driver Information Systems* na área de "HORIZONTAL RTD ACTIVITIES - Language Engineering" do programa "TAP"
   Data: de Novembro de 95 a Outubro de 97

Publicações :

Figura 5: Procura de Isabel Trancoso no busca por pessoas

Além disso, uma nova versão do Busca foi implementada utilizando subjacente o sistema de codificação de corpora IMS-CWB, para procuras em estilo AC/DC.

Menuseador: Sistema de gestão do portal

O elevado número de ligações manuseadas nos vários catálogos, associado à frequente reordenação e recategorização destes, levou à concepção de um sistema de armazenamento plano e à criação de vários programas que construíssem a estrutura do sítio a partir das categorias atribuídas aos recursos.

Além disso, estes programas de gestão dos catálogos também efectuam a contagem automática dos recursos em cada categoria, assim como impõem um estilo normalizado e garantem a datação apropriada, tornando a manutenção muito mais simples e flexível.

Eixo 2: Disponibilização e criação de recursos

Um dos objectivos da Linguateca como centro de recursos é o de aumentar significativamente o número e a qualidade dos recursos públicos para o processamento da língua portuguesa.

Esse objectivo foi plenamente alcançado, existindo agora mais recursos para o português (escrito) do que para a maior parte das línguas, não nos parecendo descabido afirmar que tal se deve em grande parte graças à intervenção da Linguateca e dos seus projectos específicos neste campo.

Dividimos a presente secção do relatório em subsecções relacionadas com cada **recurso** "maior"

- AC/DC
- CETEMPúblico e CETENFolha
- COMPARA
- Floresta Sintá(c)tica

e dedicamos outras tantas subsecções a **serviços** informáticos relacionados com a disponibilização de recursos, alguns deles directamente associados com os recursos acima

- AnELL
- Corpógrafo (Gestor de corpora)
- TrAva: obtenção de julgamentos sobre traduções automáticas
- DISPARA
- Águia

Também apresentamos o trabalho associado à obtenção de estatísticas e informação sobre o uso destes serviços.

Além disso, indicamos brevemente outros recursos ou iniciativas levadas a cabo peloa Linguateca no âmbito do eixo 2, tais como

- criação de corpora de raiz
- instituição de um repositório
- programas de observação do uso dos serviços
- mini-serviços
- programas desenvolvidos

Projecto AC/DC, www.linguateca.pt/ACDC/

O objectivo inicial do projecto AC/DC era colocar todos os corpora existentes para os quais obtivemos autorização para dar acesso acessíveis de um ponto só na rede. Além disso, os corpora foram melhorados com informação linguística de dois tipos:

1. segmentação em frases, parágrafos, e unidades textuais
2. classificação gramatical e análise sintáctica (com a colaboração de Eckhard Bick, usando o seu analisador sintáctico PALAVRAS[6])

A tabela 3 indica o tamanho e constituição dos ditos, a 1 de Setembro de 2003.

Tabela 3: Apanhado dos corpora do projecto AC/DC (a ordem é arbitrária)

Nome	**Unidades**	**Palavras**	**Frases**	**Breve descrição**
FRASESPP	19.340	16.225	594	Frases em português de Portugal
... anotado	19.542	16.208	594	
FRASESPB	22.486	19.155	651	Frases em português do Brasil
... anotado	22.730	19.165	651	
CPPRMI	1.198.015	997.695	38.151	Texto jornalístico, Público, Portugal, 1991-98, dois parágrafos / extracto
... anotado	1.202.938	995.851	38.251	
ANCIB	811.739	650.045	25.798	Correio electrónico, lista brasileira ANCIB, 1998-2003
... anotado	828.475	660.045	25.596	
DIACLAV	7.441.109	6.488.273	228.856	Diários regionais Diário de Coimbra, Diário de Leiria, Diário de Aveiro, Viseu Diário, Portugal, 1999-2000
... anotado	7.529.495	6.549.823	210.741	
AVANTE	7.607.651	6.488.201	204.686	Semanário partidário, Avante!, Portugal, 1997-2002
... anotado	7.685.242	6.512.510	204.833	
NATURA	7.257.175	6.257.950	225.673	Texto jornalístico, Público, Portugal, 1991-1994, dois parágrafos / edição
... anotado	7.321.642	6.268.817	225.734	
ENPCPUB	89.864	72.244	4.369	Literatura traduzida do inglês, de 5 obras do ENPC
... anotado	90.574	72.392	4.369	
MINHO	2.083.761	1.738.475	53.040	Artigos de jornal regional, Diário do Minho, Portugal, antes da revisão
... anotado	2.107.826	1.747.274	53.185	
ECI-EBR	891.687	722.012	45.530	Texto brasileiro, do corpus Borba-Ramsey, compilado pelo ECI
... anotado	898.542	723.007	44.689	
ECI-EE	30.157	26.515	780	Chamada do programa europeu ESPRIT, em português de Portugal
... anotado	31.127	27.140	780	
SAOCARLOS	41.372.756	32.091.996	1.955.166	Corpus NILC, português brasileiro, texto jornalístico, cartas comerciais e texto didáctico e literário
... anotado	41.917.324	32.378.253	1.955.340	
CLASSLPPE	1.872.381	1.307.334	74.174	Texto literário Clássicos da Língua Portuguesa, Porto Editora
AMOSTRA	124.655	98.444	4.925	Amostra do corpus NILC
... anotado	124.836	98.505	4.965	
CETEMPúblico	229.038.019	191.687.833	7.082.094	Texto jornalístico, Público, Portugal, 1991-98, dois parágrafos / extracto
... anotado	229.861.093	190.935.437	7.081.564	

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

Durante o desenvolvimento do projecto, tornou-se também evidente a vantagem de criar outros corpora de raiz, e, de facto, neste momento os corpora servidos pelo AC/DC foram maioritariamente criados na Linguateca.

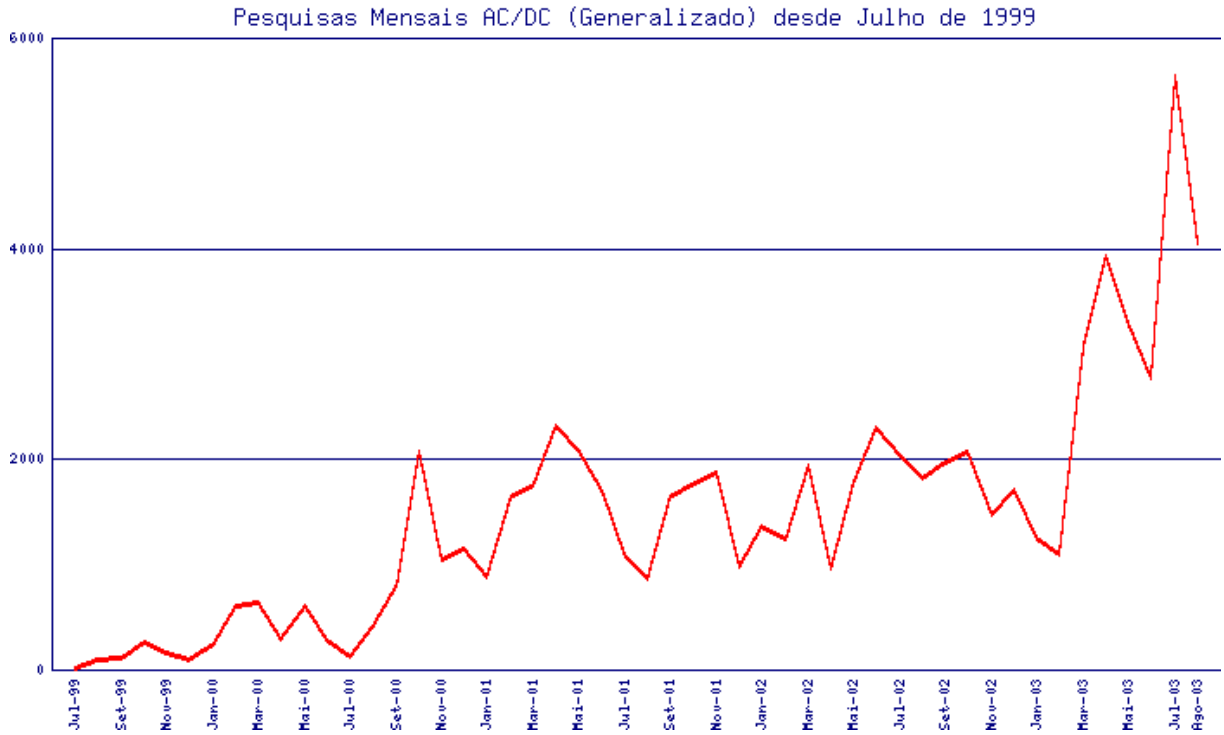


Figura 6: Uso mensal dos serviços do projecto AC/DC (incluindo Floresta)

Uma breve descrição da utilização deste projecto é dada nos gráficos de pesquisas mensais (Figura 6, acima) e de distribuição geográfica dos utilizadores (Figura 7, abaixo).

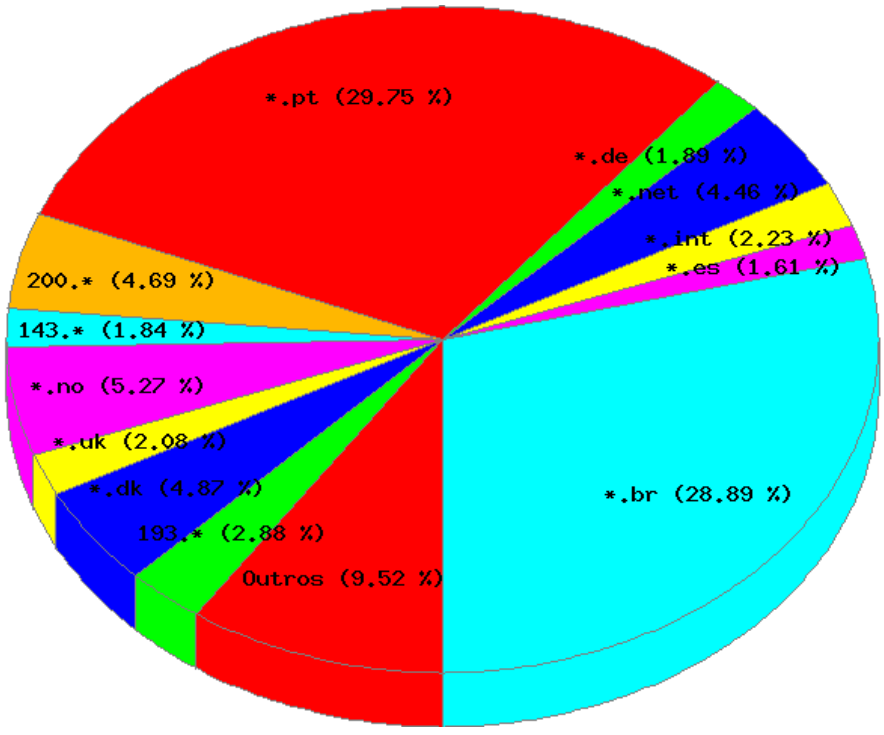


Figura 7: Distribuição geográfica dos clientes do serviço AC/DC

Note-se que o projecto AC/DC engloba vários serviços distintos, conforme o tipo de problemas que o utilizador quiser resolver, e que o projecto Floresta Sintá(c)tica (ver pontos seguintes) se constituiu como uma extensão do projecto AC/DC para associar mais informação linguística de forma a permitir resolver problemas ainda mais complicados.

De momento, existem três serviços:

1. comparação de corpora: frequências e distribuição conjunta de uma palavra em relação a um determinado atributo, em todos os corpora
2. frequência e ordenação de palavras (formas ou lemas) ou expressões regulares, em cada corpus
3. concordâncias e distribuição, em cada corpus

Além disso, convém realçar que damos apoio a todos os utilizadores que nos perguntam ou põem problemas, assim como criámos extensiva documentação, e capacidades de documentação automática para ter sempre a última informação sobre os corpora a que o utilizador acede.

Do ponto de vista técnico, desenvolvemos e aprimorámos no AC/DC uma biblioteca de processamento de corpora portugueses, que atomiza e separa em frases, tendo em conta uma grande quantidade de abreviaturas e outras convenções gráficas para não-palavras.

Além disso, desenvolvemos para cada corpus filtros que permitem interpretar as convenções de cada género textual e identificar características especiais, como campos em mensagens electrónicas, autores em notícias de jornal, ou falas de personagens de peças de teatro, para citar apenas alguns.

A figura 8 dá uma ideia do processamento técnico envolvido na disponibilização dos corpora na rede.

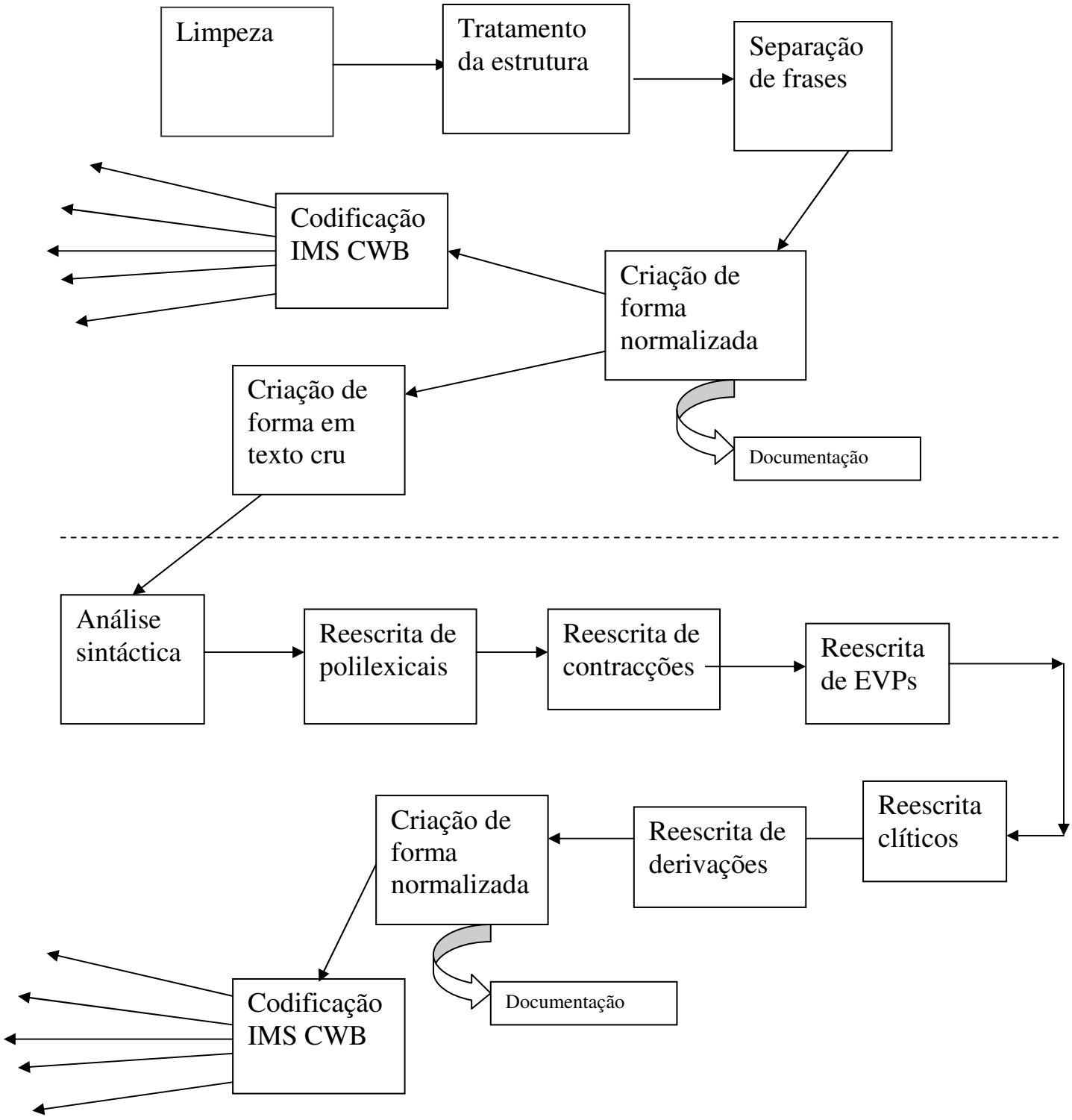


Figura 8: Processamento técnico envolvido nas duas fases do AC/DC (corpora crus, corpora anotados)

Projectos CETEMPúblico e CETENFolha

Além de acessíveis através do projecto AC/DC, são também passíveis de obtenção na sua totalidade dois grandes corpora de linguagem jornalística em português, separada em extractos de dois parágrafos, reordenados de forma a não permitir a reconstrução das notícias completas.

www.linguateca.pt/cetempublico/

A versão 1.0 do CETEMPúblico (**C**orpus de **E**xtractos de **T**extos **E**lectrónicos **MCT/Público**) foi criada a 25 de Julho de 2000 e contém cerca de 190 milhões de palavras distribuídas por um milhão e meio de extractos, correspondentes a cerca de 1500 edições diárias do jornal *PÚBLICO* (algumas delas incompletas), quase inteiramente em português europeu.

A obtenção da autorização do PÚBLICO foi conseguida por meio de um protocolo entre o jornal e o Ministério da Ciência e da Tecnologia em 2000.

O CETEMPúblico tem uma página dedicada onde é publicada informação actualizada sobre o corpus. Este, além de distribuído gratuitamente em CD, organizado em vinte volumes, encontra-se também acessível através do projecto AC/DC. Uma breve descrição quantitativa deste recurso, na sua versão 1.7 (até agora o maior corpus compilado para a língua portuguesa), encontra-se na tabela seguinte:

Tabela 4: Descrição quantitativa do CETEMPúblico

Característica	Tamanho
Unidades	229.038.019
Unidades diferentes	1.033.041
Palavras	191.687.833
Palavras diferentes	999.059
Extractos	1.504.258
Parágrafos	2.571.735
Frases	7.082.094
Títulos	655.059
Autores	247.392
Elementos de lista	80.060

A versão 1.7 foi disponibilizada também através do LDC, Linguistic Data Consortium, de forma a garantir uma existência mais longa ao recurso, e também para lhe dar um controlo de qualidade internacional.

O número de investigadores ou grupos de investigaçãoque nos pediram directamente o CETEMPúblico (em CD gratuito) foi, até ao momento, de 257. De notar que não contamos com os investigadores que usam o CETEMPúblico através do projecto AC/DC, nem com aqueles que acedem ao CETEMPúblico através de outros serviços de acesso a corpora, como foi o caso do sistema de Mark Davies da Universidade de Illinois, EUA. Também não contabilizamos aqui as cópias distribuídas através do LDC.

www.linguateca.pt/cetenfolha/

O CETENFolha (**C**orpus de **E**xtractos de **T**extos **E**lectrónicos **N**ILC/**Folha** de São Paulo) é o "irmão mais novo" do CETEMPúblico e seguiu a mesma forma de criação. O texto nele incluído é o do ano de 1994 da *Folha de São Paulo*, que já fazia parte do corpus do NILC/São

Carlos. Agradecemos pois a colaboração do Núcleo Interinstitucional de Lingüística Computacional (NILC) na disponibilização do corpus.

Por corresponder a um volume de texto bastante menor, a sua disponibilização foi feita apenas através de FTP (file transfer protocol). O CETENFolha foi anunciado em 25 de Setembro de 2002 e já 103 pessoas ou grupos de investigação o encomendaram e obtiveram.

O CETENFolha é também distribuído com a sua versão anotada pelo PALAVRAS.

A tabela 5 apresenta uma breve caracterização deste recurso. "Situação" indica a informação sobre a localização do repórter; "caixa" descreve uma espécie de resumo no início do texto da notícia.

Tabela 5: Descrição quantitativa do CETENFolha

Característica	**Tamanho**
Unidades	33.247.929
Unidades diferentes	357.759
Palavras	25 .475.272
Palavras diferentes	343.620
Extractos	34.094
Parágrafos	688.400
Frases	1.597.807
Títulos	147.238
Autores	80.133
Elementos de lista	49.721
Caixas	20.407
Situações	4 470

A cada um destes recursos está associada uma lista de discussão electrónica, gerida pela Linguateca, nomeadamente cetempublico@linguateca.pt e cetenfolha@linguateca.pt.

Projecto COMPARA/DISPARA, www.linguateca.pt/COMPARA/

O projecto COMPARA/DISPARA, lançado em colaboração com Ana Frankenberg-Garcia, tem como objectivo criar e disponibilizar o acesso a um corpus paralelo aberto de traduções entre português e inglês e vice-versa, permitindo grande número de opções de procura e uma interface amigável que sirva vários tipos de utilizadores e várias necessidades de informação contrastiva diferentes.

A primeira versão do corpus foi tornada pública em Agosto de 2000, com 4 pares de textos. Na hora da escrita do presente relatório, o corpus encontra-se na sua versão 4.0, com o conteúdo que se apresenta nas tabelas 6 a 8.

Tabela 6: Número de pares de textos incluídos na versão 4.0 do COMPARA

Textos alinhados	Originais	Traduções
Português	21	13
Inglês	11	22
Total	32	35

Tabela 7: O tamanho do COMPARA em número de palavras e de unidades de alinhamento, na versão 4.0 do COMPARA

Palavras	Originais	Traduções	Originais & Traduções	Unidades de alinhamento
Português	292947	360115	653066	20070
Inglês	360603	321918	682521	22014
Total (2 línguas)	653550	682033	1335587	42084

Tabela 8: Marcação de informação relevante na versão 4.0 do COMPARA

Língua	Notas de tradução	Palavras ou expressões estrangeiras	Nome de entidades	Títulos	Palavras ou expressões com ênfase
Português	50	1084	199	452	424
Inglês	91	614	69	451	591
Total	141	1698	268	903	1015

Estamos convencidos de que o COMPARA é, neste momento, o maior corpus paralelo editado que contém o português. De facto, pensamos que, dos corpora paralelos editados e revistos manualmente, é o maior existente na actualidade. O corpus continua em franca expansão, com mais 21 textos em fila de espera já com as autorizações concedidas.

Além disso, o sistema de desenvolvimento do corpus, com um fluxograma complicado de revisão e controlo de qualidade, junto com o processamento técnico associado, que originou no ambiente do projecto AC/DC mas que se foi progressivamente independentizando, foi chamado de sistema DISPARA. O DISPARA começou por ser o processamento técnico associado ao COMPARA e acabou por ser concebido como um sistema genérico de disponibilização de corpora paralelos sobre o sistema de codificação IMS-CWB, que pode ser usado com outros corpora paralelos, com outras línguas e funcionalidades.

Desde a sua criação, e precisamente devido ao seu conteúdo multilingue (e ao facto da interface na rede ser estritamente paralela também, com documentação sempre em português e inglês), o COMPARA teve uma base considerável de utilizadores internacional, como o

prova a figura 10, assim como uma utilização que se estabilizou na ordem das 1200 consultas por mês.

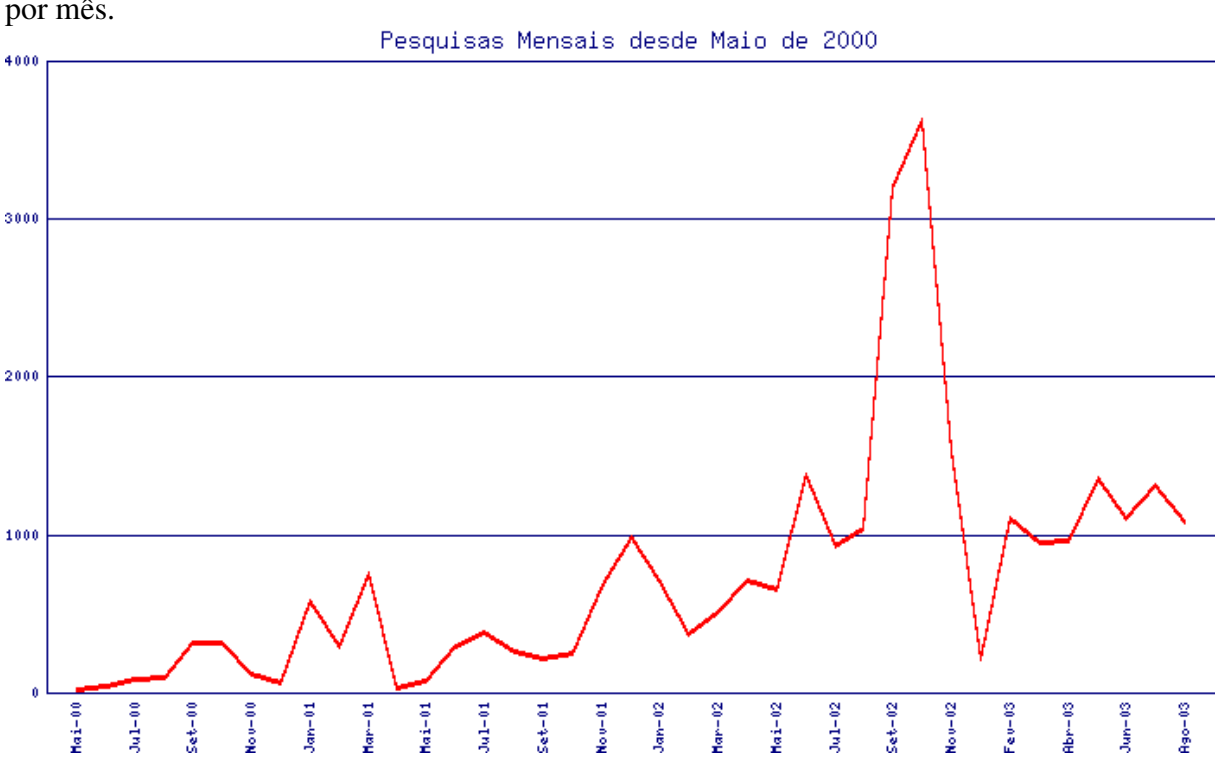


Figura 9: Consultas ao COMPARA desde a sua criação

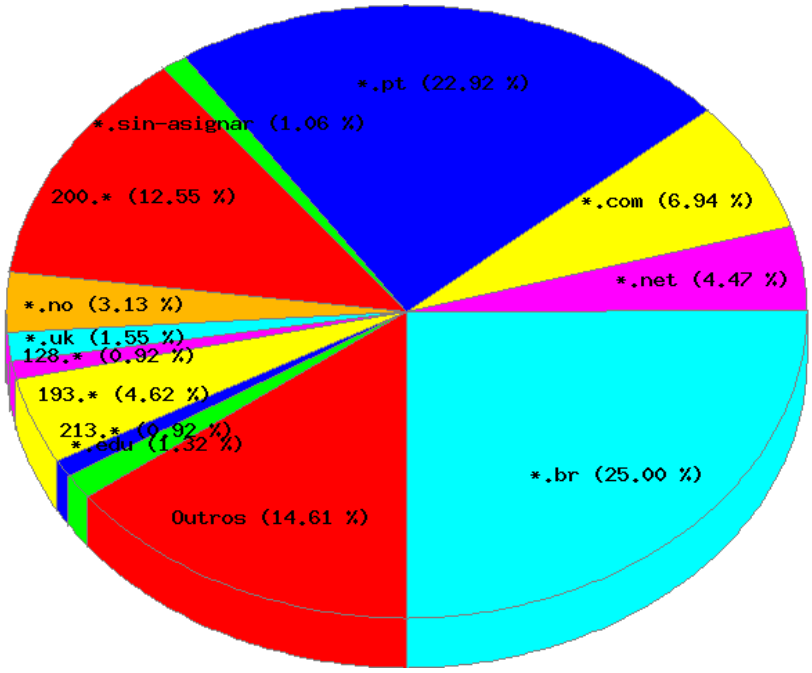


Figura 10: Distribuição dos utilizadores do COMPARA

Projecto Floresta Sintá(c)tica, www.linguateca.pt/Floresta/

O projecto Floresta Sintá(c)tica foi lançado em 1999 em colaboração com Eckhard Bick para obter um *treebank* para o português, como uma sequela (ou melhoria) do projecto AC/DC, que disponibilizava corpora anotados automaticamente sem revisão humana.

Usando o PALAVRAS, o analisador sintáctico do projecto VISL e o primeiro milhão inicial dos corpora CETEMPúblico e CETENFolha, uma equipa de linguistas foi contratada para rever e documentar tanto o formato dependencial CG produzido pelo PALAVRAS como as árvores de estrutura sintagmática (em inglês, "phrase structure") que eram automaticamente calculadas a partir desse formato.

O projecto da Floresta Sintá(c)tica, devido ao grande número de intervenientes e ao facto de ser um projecto inovador, teve um rendimento relativamente baixo (em termos de número de árvores), mas serviu para produzir um recurso único para a língua portuguesa e para formar várias pessoas na área da sintaxe computacional.

De momento existem 1.596 árvores revistas, correspondendo a 281 extractos do CETEMPúblico (1.561 frases distintas), que totalizam 41.115 unidades, e dizem respeito a aproximadamente 39 mil palavras. Além disso, também disponibilizamos todo o primeiro milhão do CETEMPúblico em forma de árvores (ainda sem revisão), a que chamamos a Floresta Virgem (41.406 árvores, não revistas).

Do ponto de vista informático, no âmbito da Floresta foram criados o Pica-pau e o Águia. O primeiro é um sistema de ajuda à revisão das árvores (em EMACS), o segundo é um sistema de procura inovador sobre o conteúdo da Floresta.

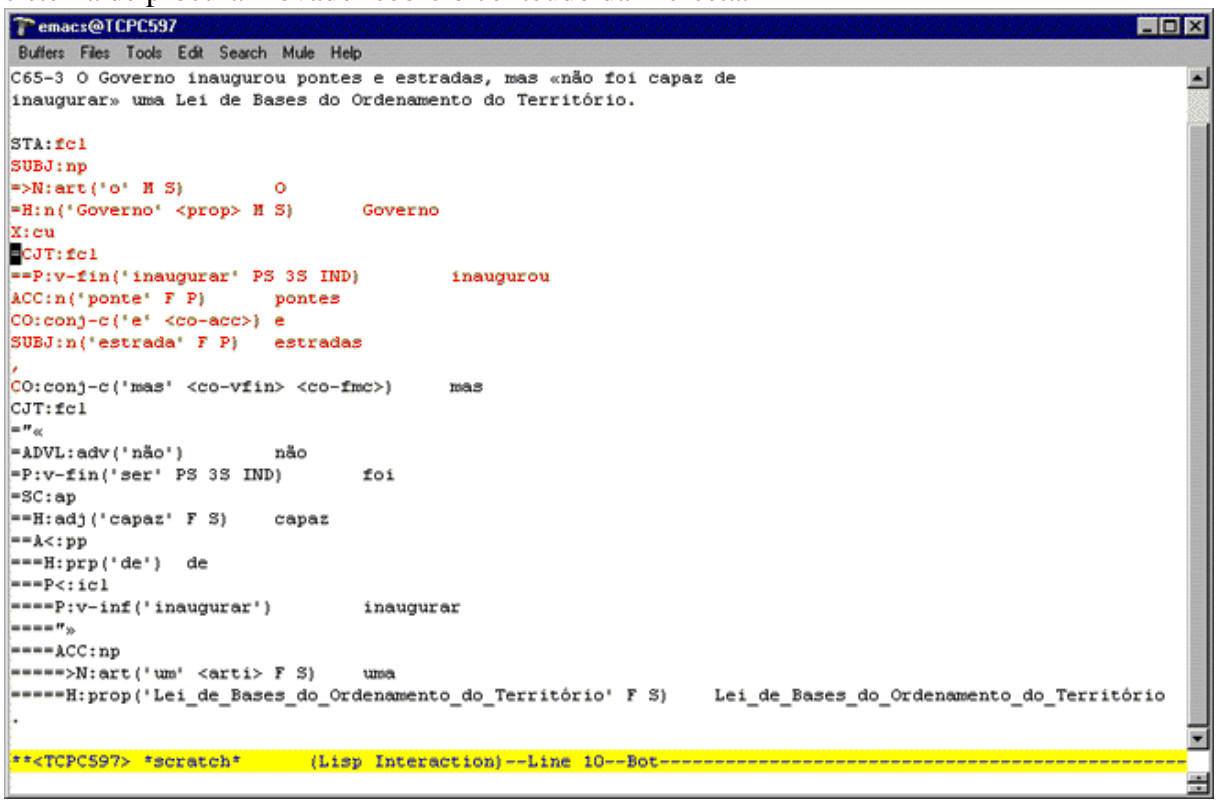


Figura 11: Exemplo de uso do Pica-pau, na terceira frase do extracto 65 do CETEMPúblico

AnELL

O AnELL, **An**otador **E**lectrónico **L**abEL-**L**inguateca, é um serviço de anotação morfossintáctica para textos não públicos, usando os recursos lexicais criados pelo LabEL[7] e o sistema INTEX[8] como sistema de processamento subjacente.

A motivação para a criação deste recurso é a de i) fornecer anotação gramatical de textos em português a uma comunidade diversificada de utilizadores interessados na língua portuguesa; ii) dar resposta ao problema dos investigadores que, tendo corpora que não podem tornar públicos, necessitam de obter informação linguística sobre o seu conteúdo; iii) criar um conjunto de filtros de pré e pós-processamento que permitam executar a anotação para o maior número possível de formatos e de utilizações, os quais deverão facilitar a sua reutilização em futuros pedidos.

Desde o início que o sistema foi pensado de forma a poder incorporar intervenção humana, na forma de actualização lexical, remoção de ambiguidades e verificação (de um subconjunto) da anotação, de forma a ser possível ajuizar a qualidade da anotação resultante. Este serviço tem, pois, dois modos: modo não supervisionado, e modo supervisionado.

Além disso, o utilizador pode escolher vários formatos de saída dos resultados, como ilustrado na figura 12. Está também previsto para breve um sistema de filtros que permitam aceitar corpora já anotados, ou com marcações estruturais complicadas, que sejam preservadas na saída do AnELL.

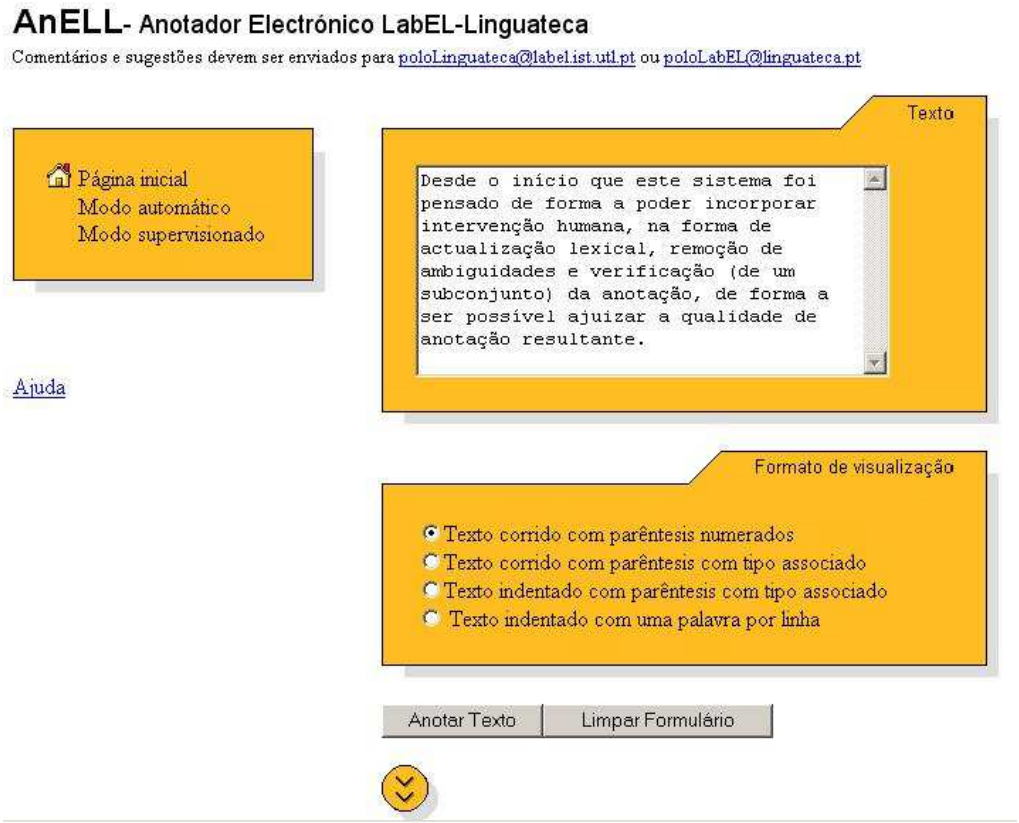


Figura 12: Formulário de entrada do AnELL

[7] Para estes recursos, consulte-se a documentação em http://label.ist.utl.pt/recursos-publicos.html.
[8] Veja-se Silberztein, Max. *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Masson Ed.: Paris, 1993

Corpógrafo: Gestor de corpora para ensino de tradução e terminologia

A principal actividade do pólo CLUP/FLUP do Porto tem sido a de criar um serviço de manipulação, criação e trabalho com corpora de domínios específicos, associada a sua formação à actividade lectiva da Professora Belinda Maia, que dirige o Mestrado em Terminologia e Tradução na Faculdade de Letras da Universidade do Porto.

Esse serviço, inicialmente chamado Gestor de Corpora (GC), mas recentemente rebaptizado de *Corpógrafo*, tem como objectivo agilizar a produção de corpora de domínios específicos, do tipo corpus "faça-você-mesmo"[9], de forma a produzir trabalho terminológico e, mais tarde, contrastivo.

É um ambiente, na rede, do tipo caixa de ferramentas, que permite, à data de escrita do presente relatório:

1. a manipulação de ficheiros dos tipos mais frequentes (quer localizados no computador do utilizador quer na rede, através da especificação de um URL): PDF, RTF, HTML etc.
2. o processamento simples desses mesmos ficheiros (fraseador, editor)
3. a escolha de um subconjunto dos ficheiros do utilizador para criação de um corpus específico
4. a detecção automática de candidatos a termos
5. a selecção e inclusão em bases de dados terminológicas
6. a edição dessas bases, adicionando informação sobre relações semânticas tais como sinonímia, antonímia, hiperonímia, etc.
7. a consulta à terminologia em contexto

Muitas outras funcionalidades estão pensadas e serão implementadas num futuro próximo, sobretudo de forma a trabalho contrastivo poder ser efectuado sobre corpora comparáveis.

Além disso, o Corpógrafo tem associado um sistema de gestão dos utilizadores (de forma a proteger os corpora individuais) e poder monitorizar e mesmo criar grupos que manipulem ou aumentem um mesmo corpus.

Este sistema, embora primordialmente usado e testado pelos utilizadores locais da Faculdade de Letras da Universidade do Porto (já conta com cerca de vinte utilizadores diferentes), encontra-se disponível para qualquer investigador interessado em corpora específicos, desde que trabalhe com o português (como uma das línguas, no caso comparável), no endereço http://www.linguateca.pt/corpografo/.

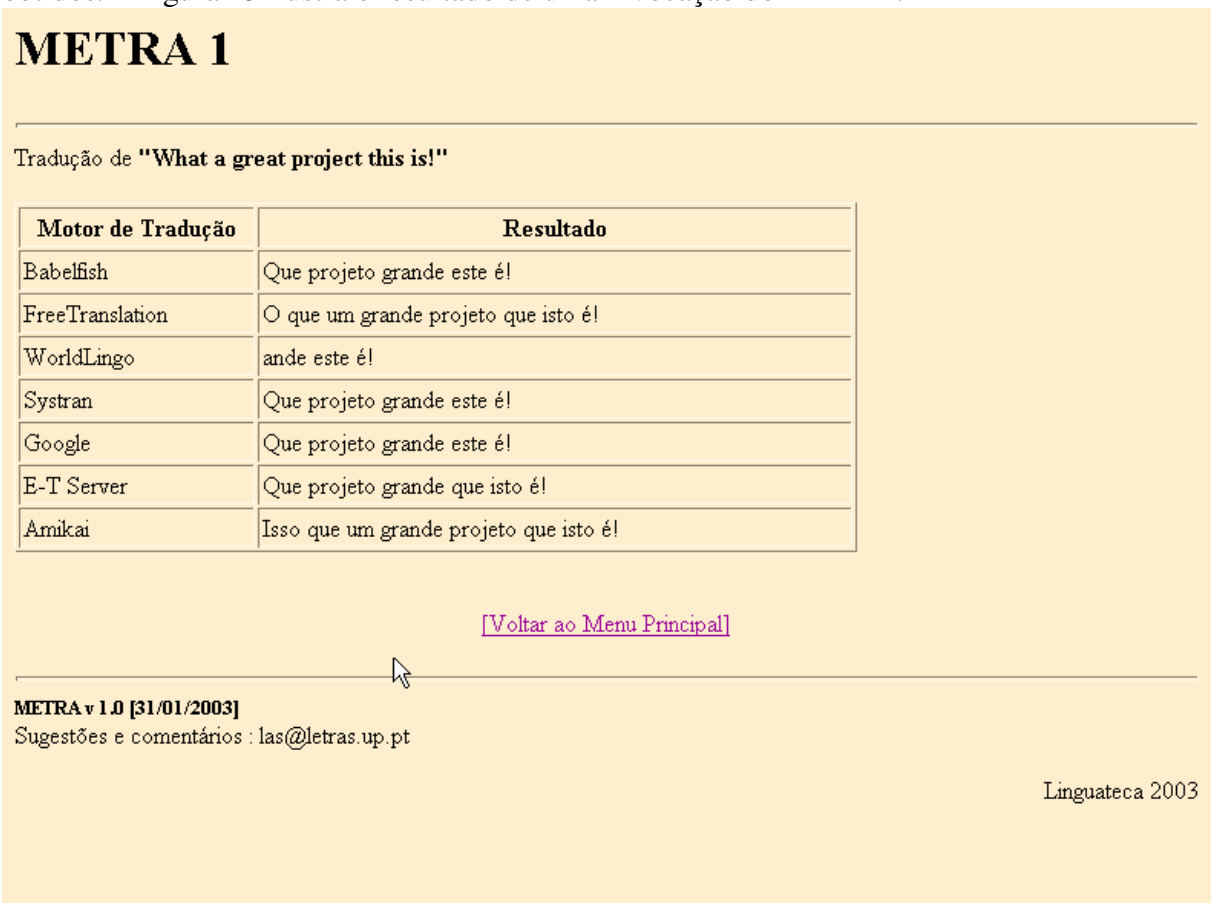
[9] Veja-se Zanettin, Federico. "DIY corpora: the WWW and the translator", in Belinda Maia, Johann Haller & Margherita Ulrych (eds.), *Training the language services provider for the New Millenium, Proceedings of the III Encontros de Tradução de Astra-FLUP*, Porto: FLUP, 2002, pp. 239-48.

Serviços relacionados com a avaliação de tradução automática

Associado às actividades de avaliação conjunta (eixo 3), também foram criados pelo pólo CLUP/FLUP do Porto alguns recursos e serviços de compilação de recursos, que passamos a descrever.

METRA: MEtaTRAdutor automático

O METRA é um serviço que submete um excerto de texto a vários motores de tradução disponíveis na rede (no momento presente, sete) e apresenta conjuntamente os resultados obtidos. A figura 13 ilustra o resultado de uma invocação do METRA.



METRA 1

Tradução de **"What a great project this is!"**

Motor de Tradução	Resultado
Babelfish	Que projeto grande este é!
FreeTranslation	O que um grande projeto que isto é!
WorldLingo	ande este é!
Systran	Que projeto grande este é!
Google	Que projeto grande este é!
E-T Server	Que projeto grande que isto é!
Amikai	Isso que um grande projeto que isto é!

[Voltar ao Menu Principal]

METRA v 1.0 [31/01/2003]
Sugestões e comentários : las@letras.up.pt

Linguateca 2003

Figura 13. Invocando o METRA

Boomerang: Determinação de Pontos de Fixação em Motores de Tradução

O BOOMERANG é um sistema que faz bumerangue de traduções inglês-português-inglês-português-inglês etc. até não se verificar mudança no resultado final. Da página http://poloclup.linguateca.pt/ferramentas/boomerang/ podemos ler : "O BOOMERANG é um serviço que submete um excerto de texto a vários motores de tradução e requisita traduções entre Português <-> Inglês até encontrar um ponto de fixação. Por ponto de fixação entende-se uma situação em que um excerto de texto se mantém igual depois de ter sido traduzido para outra língua e de volta para a língua original."

A figura 14 ilustra o funcionamento do sistema.

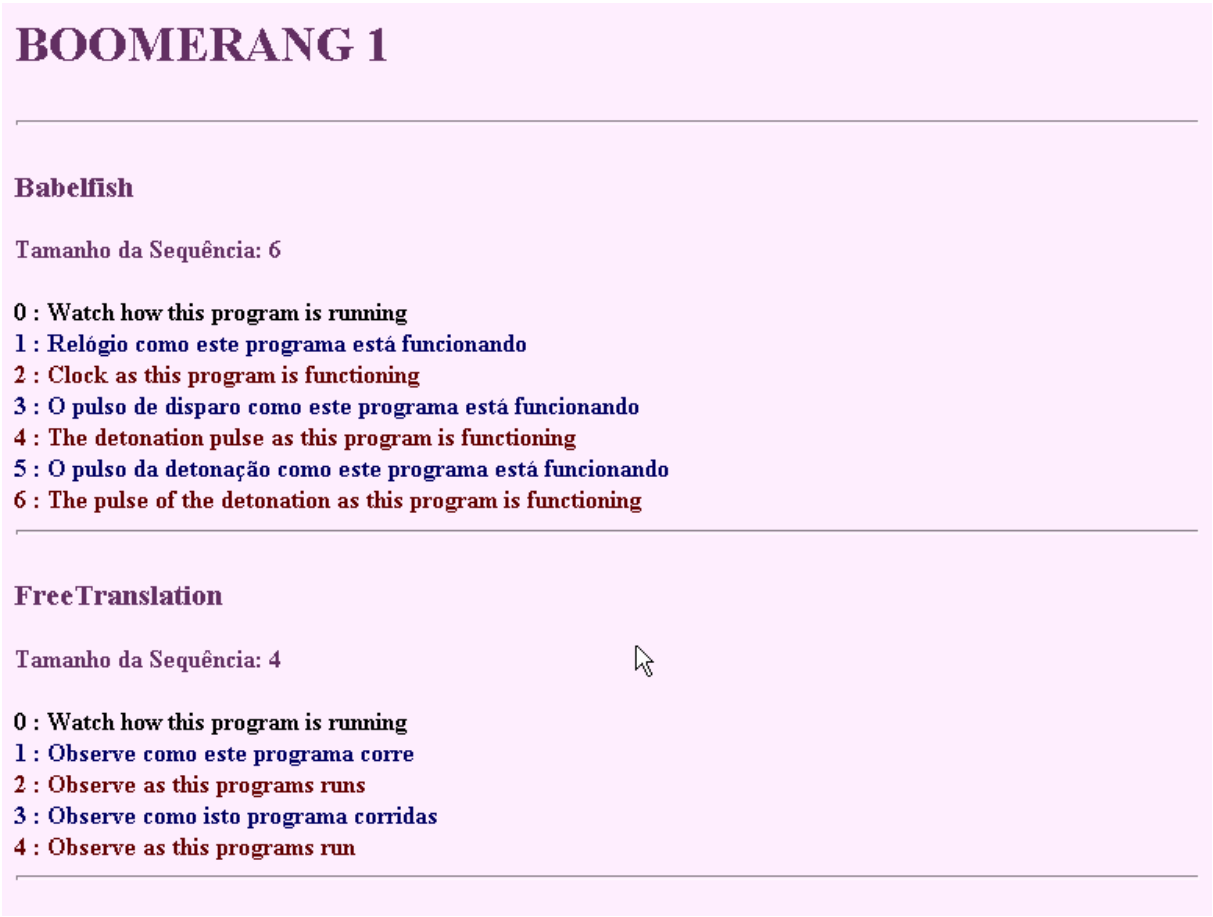


Figura 14: Usando o Boomerang

TrAva: Compilação de julgamentos de qualidade sobre traduções automáticas

O TrAva (**Tra**dução e **Ava**liação) é um serviço de compilação de julgamentos de qualidade da tradução automática, que permite guardar, num formato comum e com um conjunto de parâmetros pré-definidos, problemas de tradução e sua classificação, para posterior análise e estudo.

A forma de o utilizar é a seguinte: O utilizador escolhe um assunto que quer testar, procura frases autênticas em inglês (de preferência no *British National Corpus*) e classifica-as em termos de padrões de sequência de classificações gramaticais. O TrAva apresenta então quatro traduções automáticas, associadas a uma grelha de classificação dos possíveis problemas, e pede ao utilizador para marcar as traduções problemáticas e classificar, se possível, o problema.

Como relataremos na secção seguinte, o uso do TrAva redundou num corpus de traduções automáticas avaliadas, o CorTA, que se encontra em permanente expansão e permite a consulta por tipo de problemas, por texto na tradução ou no original, por motor de busca, etc.

DISPARA

O DISPARA é um sistema de disponibilização de corpora paralelos na Web que foi desenvolvido para o COMPARA, mas que se autonomizou e pode ser considerado um sistema genérico de criação e serviço de corpora paralelos.

Tem três ingredientes principais:

1. Em primeiro lugar, um sistema de processamento (”workflow”) complicado, entremeando revisão humana em três fases.
2. Depois, um conjunto de ferramentas para codificar e automatizar a codificação dos corpora, associados ao IMS-CWB, de forma a reutilizar o mais possível o trabalho feito para os corpora portugueses, também para o inglês.
3. E uma interface na Web (que depende, claro, do corpus que serve, mas que tem várias funcionalidades originais), e um serviço de documentação automática do conteúdo do corpus.

Em relação à interface, convém indicar que é absolutamente paralela nas duas línguas, ou seja, as mensagens de erro, a apresentação dos resultados, as páginas de procura, etc. existem paralelamente (e consistentemente) nas duas línguas em que o COMPARA pode ser interrogado – e poderiam ser criadas em quantas línguas se quisesse.

Além disso, separa claramente as duas funções ”procura” e ”mostra”, e permite algumas procuras que, ao que sabemos, não existem noutros sistemas de corpora, tal como a distribuição cruzada, a procura por reordenamento ou por notas de tradução. Na figura 15, mostramos um extracto do resultado da procura por unidades de tradução reordenadas.

Concordância

Procura: **show +place; show +reord; <ua> [_.texto="P.*" & (_.ua=".*tipo=1-[1-9/+]+re.*")] expand to ur;**.

PAJA1(72):	**Correia Balduíno, por exemplo, que depois de velho dera em ateísta. O comerciante torcia-se de riso de cada vez que João Maria entrava bêbado na vila, levantado a voz em novas blasfémias.**	Correia Balduino, for example. This merchant, ***who with age had become an atheist*** , would bend over with laughter every time João Maria entered drunk into town, raising his voice in new blasphemies.
PBPC1(460):	**O rapaz ficou impressionado com o que viu, e lembrou-se do brilho que havia notado no dia anterior. O velho tinha um peitoral de ouro maciço, coberto de pedras preciosas.**	and the boy was struck by what he saw. . The old man wore a breastplate of heavy gold, covered with precious stones. ***The boy recalled the brilliance he had noticed on the previous day***
PBPC1(652):	**"As pedras servem para adivinhação. Chamam-se Urim e Tumim ".**	. "They're called Urim and Thummim, ***and they can help you to read the omens***
PBPM1(633):	**Cachorro que não late? , ele perguntou. Quem fez isso?**	' The dog that doesn't bark? Who did that? ' ***he asked.***
PBPM1(866):	**Para: Wilmer. De: José Guber.**	From: José Guber ***To: Wilmer da Silva***
PBRF1(2559):	**Voltei a falar com Maurício. "Trinta mil dólares é suficiente."**	. ' Thirty thousand is enough, ' ***I told Maurício***
PMMC1(410):	**Afinal, esse Zuzé! Era mesmo, o gajo.**	So it was true, ***after all, about this fellow Zuzé.***
PPCP1(13):	**Assim sendo, e na sequência dos factos ocorridos no dia três de abril do corrente ano de mil novecentos e sessenta,**	Such, then, is Inspector Elias, or Graveyard, who, because an unidentified body was found -- some

Figura 15. Procurando reordenamentos no COMPARA

Águia

O Águia é um sistema de acesso na rede a texto sintacticamente analisado em forma de árvores ("treebanks", que designaremos no que se segue por florestas), desenvolvido no âmbito do projecto Floresta Sintá(c)tica, sobre o sistema IMS-CWB. A sua motivação foi a de obter um sistema de alto nível que cativasse utilizadores não familiares com o formalismo usado na Floresta, e dessa forma obter também realimentação sobre o tipo de consultas que eventualmente faria sentido perguntar a uma floresta sintáctica.

Também com funcionamento paralelo em português e inglês, é possível procurar por tipo de sintagma, por função sintáctica, além de procuras no texto (concebido como uma extensão ao projecto AC/DC).

Como marca distintiva, o seu resultado é sempre texto, ou seja, não se dá ênfase à forma de representação das árvores subjacentes. Isto porque, por um lado, se pretende minimizar a informação de que o utilizador precisa para consultar o recurso e, por outro, porque todas as árvores da Floresta já são distribuídas publicamente no seu formato interno.

A figura 16 ilustra o resultado de uma consulta não trivial através do Águia.

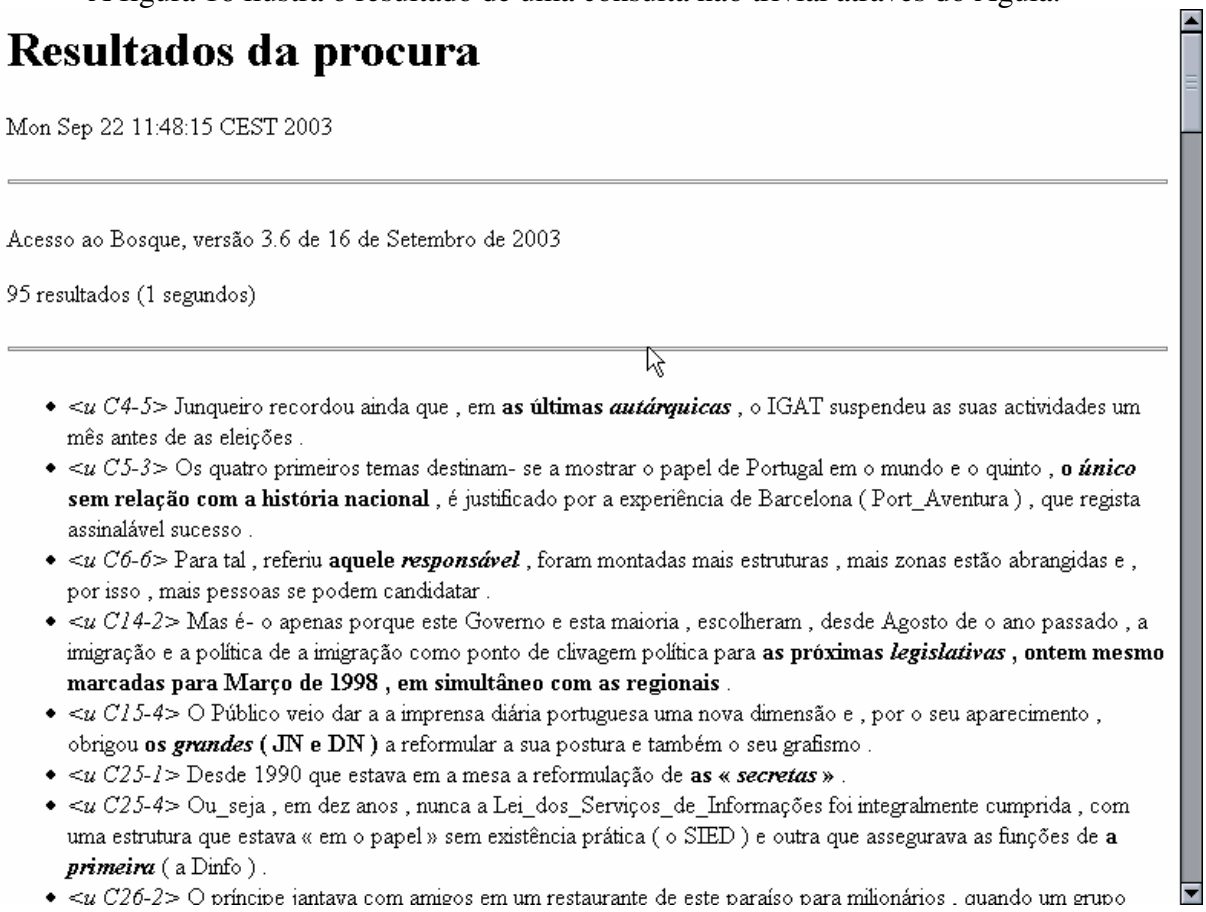


Figura 16: Procurando sintagmas nominais cujo núcleo seja um adjectivo

Corpora criados de raiz

Conforme explicado na secção relativa ao projecto AC/DC, a Linguateca acabou por criar vários corpora associados a necessidades específicas, que foram posteriormente integrados e tornados acessíveis através do AC/DC. Em alguns casos, especificados no seguimento, os corpora já existiam, mas foram reformulados e tornados públicos no âmbito da Linguateca.

Descrevem-se muito sucintamente esses vários recursos no que se segue:

ANCIB: tráfego na lista do mesmo nome da *Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (ANCIB)* brasileira. Para este corpus, foi preciso desenvolver um processamento do formato de correio electrónico armazenado por dois sistemas diferentes (Eudora e Outlook); proceder à identificação de material citado; melhorar o sistema de atomização (para dar conta dos URLs e dos programas de conferências, que abundam na lista), assim como implementar um sistema detector de língua (além do português, mais duas línguas são usadas na lista, nomeadamente o inglês e o castelhano). Este corpus encontra-se em permanente expansão, visto que a lista se mantém activa.

AmostRA-NILC: Corpus anotado com classificação gramatical, desenvolvido por Rachel Aires para efeitos da sua tese de mestrado sobre a comparação de anotadores automáticos para o português brasileiro, e que é uma amostra do Corpus NILC/São Carlos.

DiaCLAV: Corpus de texto jornalístico compilado a partir das edições na rede de quatro diários regionais portugueses, o *Diário de Coimbra*, o *Diário de Leiria*, o *Diário de Aveiro* e o *Viseu Diário*.

Avante: Corpus de texto jornalístico compilado a partir da edição na rede de um semanário partidário, o *Avante!*.

ClassLPPE: Obras clássicas da literatura portuguesa, digitalizadas e editadas pela Porto Editora. Para este corpus, foi preciso idealizar uma forma de codificar poesia e teatro.

CorTA: Corpus de traduções automáticas de inglês para português, um original e quatro traduções, a que estão associadas um manancial de informação de avaliação, compilado a partir do TrAva (ou seus predecessores EVAL1, EVAL2...) no pólo do Porto

Podemos, além disso, indicar que, à data da escrita do presente relatório, estão em fase adiantada de conclusão e/ou disponibilização os seguintes corpora adicionais (cujos nomes ainda são provisórios):

CHAT: Corpus da utilização de chat, compilado na FLUP.

AMorfo: Corpus de extractos de outros corpora, analisado morfologicamente por seis sistemas diferentes, compilado como um dos resultados das Morfolimpíadas.

CoNE: Corpus de **Co**rreio **N**ão **E**ndereçado em português.

Programas de observação do uso dos serviços

Uma tarefa associada ao desenvolvimento, manutenção e melhoria de todos os serviços desenvolvidos pela Linguateca é a observação do uso e dos problemas encontrados pelos seus utilizadores.

Para permitir uma visão global, foi criado pelo pólo do Porto uma colecção de programas que calculam, a partir da informação rotineiramente compilada por cada serviço, as seguintes estatísticas:

- Evolução do número de acessos e distribuição geográfica dos utilizadores (identificáveis) – resultados utilizados, aliás, nas figuras 6, 7, 9 e 10 do presente relatório
- Total de resultados, distribuído por tipos de pesquisas (dependendo do serviço)
- Tipos de utilizadores, por número de pesquisas, sessões, e resultados
- Tipos de sessões (definidas como um conjunto de pesquisas individuais sucessivas realizadas do mesmo computador e não distando mais de quinze minutos entre elas): por duração, número de pesquisas, tamanho dos resultados
- Origens (identificação do computador) por número de sessões e número de resultados

Este tipo de panorâmica permite ter uma visão global da nossa massa dos utilizadores e também dos tipos de procuras mais populares.

Repositório

A Linguateca organizou também um serviço de repositório para alojar recursos cujos criadores não tivessem lugar ou capacidade na rede para tal, desde artigos e obras de referência a recursos como colecções ou léxicos.

Apesar de este serviço não ser muito usado pela comunidade (de momento inclui apenas 38 publicações, uma colecção de recolha de informação (RI), um conjunto de regras de tradução automática e uma faixa de demonstração de uma base de dados etiquetada para sintetizador de fala), constituiu mais uma medida para tornar ferramentas ou dados disponíveis a uma maior audiência, e, sobretudo, manter a possibilidade de esses recursos sobreviverem aos projectos no âmbito dos quais foram criados.

Mini-serviços

No âmbito de projectos maiores, a Linguateca produziu também alguns "mini-serviços", ou seja, serviços que se podem utilizar independentemente dos projectos maiores em que foram concebidos:

- Um serviço de alinhamento de textos traduzidos de ou para o português, usando o sistema EasyAlign subjacente (autonomizado do DISPARA).
- Um serviço de extracção do texto de ficheiros de vários outros tipos, o EXTEX (autonomizado do Corpógrafo).
- Um calculador de estatísticas a partir do texto analisado pelo PALAVRAS (autonomizado do AC/DC).

Programas

Durante os vários anos de actividade, várias dezenas de programas em Perl foram desenvolvidos e partilhados entre os vários projectos e pólos da Linguateca.

Algum esforço terá de ser feito para aumentar a audiência desses programas e disponibilizar o código fonte para o público em geral. Alguns primeiros passos nesse sentido foram já dados pelo pólo de Braga ao criar módulos públicos de Perl.

Eixo 3: Avaliação conjunta

Avaliação conjunta (do inglês "evaluation contest") é um modelo de avaliação em que vários grupos comparam, com base num conjunto de tarefas consensuais, o progresso dos seus sistemas numa dada área, usando para isso um conjunto de recursos comum, e uma métrica consensual.

A motivação para actividades nesta área é simples, e corresponde, aliás, à tentativa de superação de vários dos entraves cedo identificados:

- a falta de comunicação entre os membros da comunidade
- a falta de padrões de avaliação

Desde 2001 que tem havido um crescendo de actividades nessa área, primeiro internamente, e depois tentando agregar o maior número de interessados e tornando-os possíveis participantes, através de iniciativas por correio electrónico, na rede, e em pessoa.

A Linguateca criou em Abril de 2002 uma lista electrónica dedicada ao tema avaliação conjunta, a lista avalia@linguateca.pt.

Como corolário da agitação em torno do assunto "avaliação conjunta", houve um encontro preparatório em Faro em Junho de 2002, em que se definiram grupos de interesse e se apresentaram algumas propostas iniciais, uma das quais (a avaliação de analisadores morfológicos) foi já completamente concretizada.

O modelo tradicional de uma avaliação conjunta é o apresentado na figura 17.

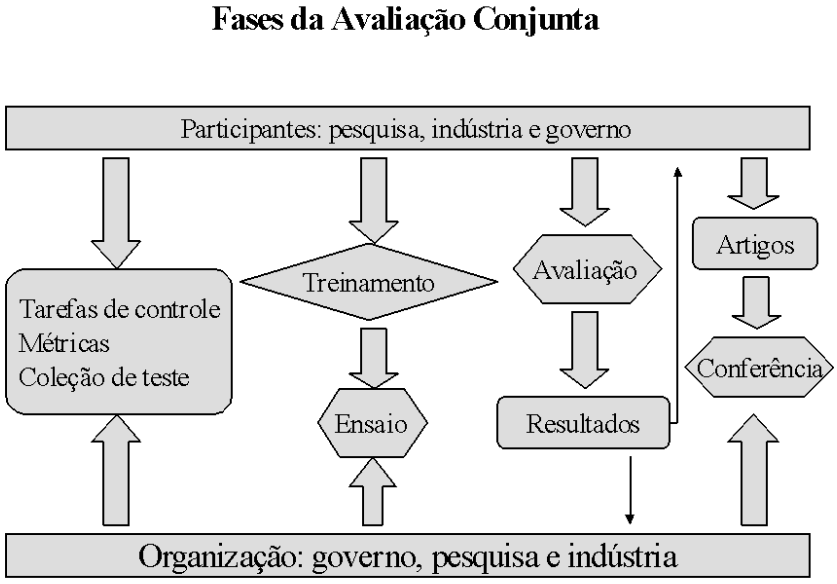


Figura 17: Avaliação conjunta: o modelo

A Linguateca desenvolveu, além da organização das *Primeiras Morfolimpíadas* para o português, várias outras actividades noutras áreas, como foi apresentado no Avalon'2003, encontro satélite do PROPOR'2003 que foi organizado pela Linguateca com a ajuda local imprescindível de Jorge Baptista.

Descrição da comunidade associada à onda de avaliação conjunta

Através da relação dos interessados na avaliação conjunta (ao todo, 113 inscritos) elegeram-se as dez áreas de (maior) interesse patentes na tabela 9.

Tabela 9: Contabilização dos interessados nas dez áreas mais votadas

Área	Inscritos	Peso
Análise morfológica	50	8,27
Léxicos monolingues	52	6,00
Recuperação de informação	34	5,68
Corpora anotados	48	5,67
Análise sintáctica	42	4,95
Extracção de informação	34	3,51
Análise semântica	31	3,43
Léxicos bilingues ou multilingues	37	3,38
Separação de frases	20	2,76
Extracção de relações semânticas	27	2,75

Outras consequências objectivas desta mentalização e popularização foram a aceitação por parte da organização do PROPOR de dedicar um dia à avaliação conjunta (o já mencionado Avalon'2003), o ter um dos dois convidados internacionais (Robert Gaizauskas) a falar de avaliação, e o facto de ter havido um aumento considerável de artigos dedicados à avaliação ou contendo secções sobre o assunto no PROPOR'2003, em comparação com as edições anteriores.

Morfolimpíadas

Após discussão na lista avalia, uma proposta inicial foi apresentada no Encontro Preparatório em Faro, em Junho de 2002, que, após algum refinamento na discussão plenária e por discussão por correio electrónico subsequente, se materializou no calendário de trabalhos da Tabela 10.

Tabela 10: Passos na organização e desenrolar das Morfolimpíadas

Passo	**Data**
Proposta inicial na lista avalia	8 de Maio 2002
Refinamento da proposta (Faro 2002)	27 de Junho 2002
Ensaio	
Anúncio (organização)	Julho 2002
Compilação cooperativa da lista dourada	até 6 de Setembro 2002
Disponibilização dos textos (organização)	10 de Setembro 2002
Envio dos resultados (participantes)	17 de Setembro 2002
Encontro presencial	1 de Outubro 2002
1.as Morfolimpíadas	
Anúncio (organização)	19 de Março 2003
Inscrição e compilação da lista dourada	até 31 de Março 2003
Disponibilização dos textos (organização)	5 de Maio 2003
Envio dos resultados (participantes)	7 de Maio 2003
Divulgação dos resultados (organização)	28 de Junho 2003
Encontro no Avalon'2003	28 de Junho 2003

Como seria de esperar, a Linguateca mantém um endereço na rede dedicado às Morfolimpíadas, http://www.linguateca.pt/Morfolimpiadas/, e estamos a preparar para

disponibilização pública tantos os textos como os resultados de todos os participantes, após anonimização destes.

Avaliação de reconhecimento de entidades mencionadas

Uma outra área em que o processo de avaliação conjunta deu alguns frutos foi a do reconhecimento de entidades mencionadas (tradução do inglês "named entity recognition"), em que o pólo do LabEL (Lisboa) se encarregou de lançar o repto, disponibilizando alguns exemplos (do CETEMPúblico e do CETENFolha) para serem marcados por vários grupos e assim se obter uma primeira panorâmica de concordância ou discordância.

Os primeiros resultados, apresentados publicamente durante o Avalon'2003, e anotados por nove diferentes participantes, apresentaram uma variabilidade descomunal (concordância de 31,6% nos dez primeiros extractos CETEMPúblico e 24,6%. nos 20 primeiros extractos do CETENFolha), indicando que era preciso definir mais rigorosamente a tarefa a comparar, assim como as métricas a utilizar.

A figura 18 fornece uma visão global da variedade de etiquetas usadas pelos vários participantes, nos 179 a 240 casos identificados.

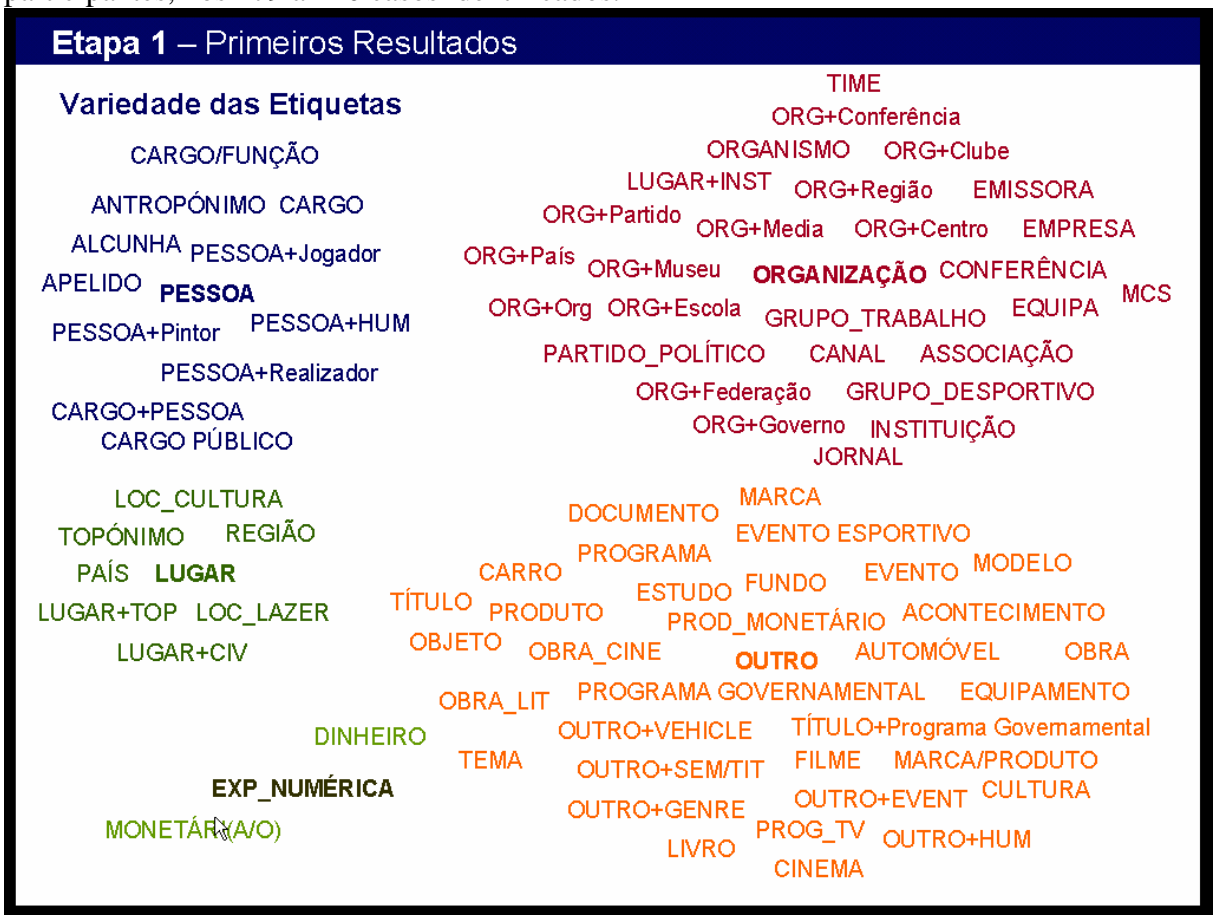


Figura 18: Variedade das etiquetas usadas pelos participantes

De momento está em preparação uma nova rodada, mais controlada, de consulta aos interessados, de forma a organizar uma verdadeira avaliação conjunta na área.

Avaliação de recolha de informação ("information retrieval")

Neste campo, a ênfase foi colocada na criação de colecções de teste que permitissem mais tarde criar avaliações conjuntas, e a Linguateca tem ajudado os donos das colecções no que respeita ao armazenamento e documentação das ditas.

Avaliação de tradução automática

Nesta área, a abordagem, liderada pelo pólo do Porto, foi completamente diferente e teve como ideia inicial a obtenção cooperativa de um conjunto de casos de teste que servissem como primeira ferramenta de avaliação.

Essa iniciativa deu origem ao METRA, Boomerang, TrAva e CorTa, já mencionados anteriormente. Neste momento pretendemos obter, durante um ano, um conjunto considerável de dados, para posterior análise (semiautomática).

Avaliação de alinhamento

A avaliação de alinhamento entre textos paralelos recaiu sob a alçada do pólo de Braga, que apresentou uma primeira proposta na lista avalia e, mais tarde, no Avalon'2003. Prevemos para breve a realização de um ensaio nesta área.

Avaliação de análise sintáctica

Nesta área não se avançou muito, mas efectuaram-se duas reflexões conjuntas, em ambos os encontros no Algarve, sobre o futuro da Floresta Sintá(c)tica como recurso de avaliação.

No entanto, uma das propostas em aberto é a de realizar uma avaliação conjunta na área da desambiguação morfológica do português como uma sequela das Morfolimpíadas, usando variados corpora (entre eles a Floresta Sintá(c)tica) como termos de comparação.

Outras actividades

Ainda que o grosso da actividade na área da Linguateca seja na promoção e organização de avaliações conjuntas, houve também algum trabalho de avaliação de recursos ou de sub-áreas do processamento computacional da língua portuguesa. Passamos a descrever brevemente quais os problemas sobre os quais nos debruçámos:

- Avaliação de corpora anotados: como medir a qualidade da anotação dos corpora do projecto AC/DC, sob várias perspectivas.
- Avaliação do CETEMPúblico: como avaliar um recurso deste tipo.
- Avaliação de conjugadores verbais: como comparar diversos sistemas que aparentemente têm o mesmo objectivo.
- Avaliação de alinhadores vs. avaliação do problema do alinhamento: como medir o problema entre o par português-inglês, e avaliar um resultado de um sistema semi-automático.
- Avaliação de ferramentas de processamento de corpora: comparação qualitativa de dois sistemas usados para o português.
- Avaliação do peso da língua portuguesa na rede: como medir o tamanho e número de páginas em português na Web.

Actividades de investigação e formação

A Linguateca não tem como objectivo realizar investigação pura ou produzir trabalho académico. Contudo, interessa-nos motivar esse trabalho quer na área da construção de recursos como na da avaliação.

Daí que os seguintes investigadores foram parcial ou totalmente apoiados pela Linguateca nos seus estudos de pós-graduação:

- **Rachel Aires**, no âmbito do seu doutoramento em Ciências de Computação e Matemática Computacional, Universidade de São Paulo, Brasil, versando uma arquitectura linguisticamente motivada para a recuperação de informação em português na rede. Esta investigadora tem sido parcialmente financiada pela Linguateca, com uma bolsa no SINTEF.
- **António Colaço**, no âmbito do seu doutoramento em Lexicografia, Universidade Nova de Lisboa, versando a exploração semi-automática de corpora, orientada para a extracção das relações sintácticas e semânticas dos seus elementos. A participação em alguns cursos e conferências deste investigador foi financiada pela Linguateca, assim como um estágio de um mês em Oslo.
- **Alberto Simões**, no âmbito do seu mestrado em Matemática e Ciências da Computação, Universidade do Minho, sobre técnicas de alinhamento de palavras em corpora bilingues. A este investigador, contratado pela Linguateca pelo pólo do Minho, foi permitido dedicar algum do seu tempo à finalização da sua tese.

Equipa associada à Linguateca

Por ordem de afectação ao projecto, a Linguateca (e/ou o Projecto Computacional do Português, que a precedeu) envolveu os colaboradores a tempo inteiro apresentados na tabela 11.

Tabela 11: Equipa da Linguateca

Nome	Data de entrada	Data de saída	Pólo
Diana Santos	15 de Maio de 1998		Oslo
Signe Oksefjell	15 de Maio de 1998	1 Agosto de 1999	Oslo
Paulo Rocha	15 de Outubro de 1999	14 de Outubro de 2000	Oslo
Renato Ribeiro Haber	1 de Outubro de 2000	30 de Setembro de 2001	Oslo
Paulo Rocha	1 de Novembro de 2000	14 de Março de 2003	Braga
Pedro Moura	1 de Outubro de 2001	31 de Agosto de 2002	Lisboa
Alexsandro Santos Soares	1 de Novembro de 2001	31 de Julho de 2002	Oslo
Cristina Mota	1 de Outubro de 2002		Lisboa
Luís Costa	26 de Outubro de 2002		Oslo
Luís Sarmento	18 de Outubro de 2002		Porto
Alberto Simões	21 de Abril de 2003		Braga

Além disso, a Linguateca envolveu os bolseiros constantes da tabela 12.

Tabela 12: Bolseiros da Linguateca

Nome	Data de entrada	Data de saída	Lugar
Rachel Aires	1 de Setembro de 2001		Oslo
Susana Afonso	1 de Setembro de 2001	30 de Junho de 2002	Odense
Ana Raquel Marchi	1 de Setembro de 2001	30 de Junho de 2002	Odense
Miguel Oliveira	1 de Setembro de 2001	28 de Fevereiro de 2002	Odense

Finalmente, os seguintes colaboradores foram contratados à tarefa. A lista é por ordem cronológica de início de actividade:

- **Tom Funcke**
- **Susana Afonso**
- **Anabela Barreiro**
- **Rosário Morais da Silva**
- **Ana Raquel Marchi**
- **Paulo Rocha**

Mais informação sobre a equipa e os perfis de cada membro encontra-se na página da Linguateca dedicada à equipa, http://www.linguateca.pt/equipa.html.

Publicações e palestras

A actividade de publicação relacionada com a Linguateca encontra-se acessível na rede, em http://www.linguateca.pt/documentos/. Nos três anos a que se refere este relatório, foram publicados ou enviados para publicação os trabalhos correspondentes aos pontos [9-36]. Parece-nos, contudo, mais natural apresentar a totalidade do material produzido no projecto desde a sua criação, para permitir uma visão de conjunto.

1. Santos, Diana. "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.linguateca.pt/branco/.
2. Santos, Diana. "Porquê processamento computacional do português e não processamento de linguagem natural?", http://www.linguateca.pt/branco/Porque.html, 24 de Março de 1999.
3. Oksefjell, Signe e Diana Santos. "Breve panorâmica dos recursos de português mencionados na Web". *Anais do Terceiro Encontro de Processamento da Língua Portuguesa (Escrita e falada), PROPOR'98* (Porto Alegre, 3-4 novembro 1998), pp. 38-47.
4. Santos, Diana e Elisabete Ranchhod. "Ambientes de processamento de corpora em português: Comparação entre dois sistemas", in *Actas do PROPOR'99* (Évora, 21-22 Setembro de 1999), pp. 257-268. [estritamente, apenas o trabalho da primeira autora foi feito no âmbito do projecto]
5. Santos, Diana. "Disponibilização de corpora através da WWW", in Palmira Marrafa & Maria Antónia Mota (orgs.), *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), Lisboa: Colibri, 1999, pp. 323-346.
6. Santos, Diana. "Comparação de corpora em português: algumas experiências", 17 de Setembro de 1999, http://www.linguateca.pt/Diana/download/CCP.ps.
7. Santos, Diana. "O computador e a tradução", in *Actas do II Seminário de Tradução Científica e Técnica em Língua Portuguesa* (Lisboa, 22 a 24 de Novembro de 1999), à espera de publicação.
8. Rocha, Paulo. Uma apreciação de diversos recursos para conjugação de verbos em português, 2 de Fevereiro de 2000, http://www.linguateca.pt/Paulo/pubs/conjug.html.
9. Santos, Diana e Eckhard Bick. "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, LREC 2000* (Athens, 31 May-2 June 2000), pp. 205-210. [estritamente, apenas o trabalho da primeira autora foi feito no âmbito do projecto]
10. Santos, Diana. "Towards language-specific applications", *Machine Translation* **14** (2), June 1999, pp. 83-112.
11. Santos, Diana e Signe Oksefjell. "Using a parallel corpus to validate independent claims", *Languages in Contrast* **2** (1), 1999, pp. 117-132.
12. Santos, Diana e Signe Oksefjell. "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.

13. Santos, Diana. "Introdução ao processamento de linguagem natural através das aplicações". Elisabete Ranchhod (ed.), *Tratamento das Línguas por Computador. Uma introdução à linguística computacional e suas aplicações*, Lisboa: Caminho, 2000, pp. 229-259.
14. Rocha, Paulo Alexandre e Diana Santos. "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), pp.131-140.
15. Santos, Diana. "O projecto Processamento Computacional do Português: Balanço e perspectivas", *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), pp. 105-113.
16. Frankenberg-Garcia, Ana e Diana Santos. "Introducing COMPARA, the Portuguese-English parallel translation corpus", in Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds.), *Corpora in Translation Education*, Manchester: St. Jerome Publishing, 2003. [estritamente, apenas o trabalho da segunda autora foi feito no âmbito do projecto]
17. Veiga, Pedro e 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 (Actas do colóquio de 25 2 26 de Janeiro de 2001)*, Lisboa: Edições Colibri, pp. 103-109.
18. Santos, Diana. "Aonde vamos em relação a aonde". Actas do Simpósio 'Redescobrindo a linguagem: Pesquisa em Linguística de Corpus, 10.o InPLA (São Paulo, 14 de abril de 2000), no prelo.
19. Santos, Diana e Paulo Rocha. "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.
20. Frankenberg-Garcia, Ana e Diana Santos. "Apresentando o COMPARA, um corpus português-inglês na Web", *Cadernos de Tradução* **5,** Universidade de São Paulo, no prelo. [estritamente, apenas o trabalho da segunda autora foi feito no âmbito do projecto]
21. Afonso, Susana, Eckhard Bick, Renato Haber e Diana Santos. "Floresta sintá(c)tica: um treebank para o português", Anabela Gonçalves & Clara Nunes Correia (orgs.), *Actas do XVII Encontro da Associação Portuguesa de Linguística* (Lisboa, 2-4 de Outubro de 2001), Lisboa: APL, 2002, pp. 533-45. [estritamente, o trabalho do segundo autor não foi feito no âmbito do projecto]
22. Afonso, Susana, Eckhard Bick, Renato Haber e Diana Santos. "Floresta sintá(c)tica: primeiro ano". http://acdc.linguateca.pt/treebank/Afonsoetal2002.ps [estritamente, o trabalho do segundo autor não foi feito no âmbito do projecto]
23. Santos, Diana. "Um centro de recursos para o processamento computacional do português", *DataGramaZero - Revista de Ciência da Informação* **3** n.1 fev/02, http://www.dgz.org.br/fev02/Art_02.htm.

24. Afonso, Susana, Eckhard Bick, Renato Haber e Diana Santos. "Floresta sintá(c)tica: a treebank for Portuguese", in Manuel González Rodríguez & Carmen Paz Suárez Araujo (eds.), *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation* (Las Palmas de Gran Canaria, Spain, 29-31 May 2002), ELRA, 2002, pp. 1698-1703. [estritamente, o trabalho do segundo autor não foi feito no âmbito do projecto]

25. Santos, Diana e Caroline Gasperin. "Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation", in Manuel González Rodríguez & Carmen Paz Suárez Araujo (eds.), *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation* (Las Palmas de Gran Canaria, Spain, 29-31 May 2002), ELRA, 2002, pp. 597-604. [estritamente, apenas o trabalho da primeira autora foi feito no âmbito do projecto]

26. Santos, Diana. "DISPARA, a system for distributing parallel corpora on the Web", in Elisabete Ranchhod & Nuno J. Mamede (eds.), *Advances in Natural Language Processing (Third International Conference, PorTAL 2002, Faro, Portugal, June 2002, Proceedings)*, LNAI 2389, Springer, 2002, pp. 209-218.

27. Santos, Diana. "*Med* e *com*: um estudo contrastivo português - norueguês", *Actas do XV congresso dos romanistas escandinavos* (Universidade de Oslo, Agosto de 2002), Oslo: University of Oslo, http://www.digbib.uio.no/roman/Art/Rf-16-02-2/por/Santos.pdf.

28. Rocha, Paulo Alexandre, Alberto Manuel Simões e José João Almeida. "Cálculo de frequências para entradas de dicionários através do uso conjunto de analisadores morfológicos, taggers e corpora.", in A. Gonçalves & C.N. Correia (orgs.), *Actas do XVII Encontro da Associação Portuguesa de Linguística,* Lisboa: APL, 2002, pp.407-418. [estritamente, apenas o trabalho do primeiro autor foi feito no âmbito do projecto]

29. Aires, Rachel e Diana Santos. "Measuring the Web in Portuguese", in Brian Matthews, Bob Hopgood & Michael Wilson (eds.), *Euroweb 2002 conference* (Oxford, UK, 17-18 December 2002), pp.198-9.

30. Santos, Diana e Luís Sarmento. "O projecto AC/DC: acesso a corpora / disponibilização de corpora", *Actas do XVIII Encontro da Associação Portuguesa de Linguística* (Porto, 2-4 de Outubro de 2002), APL, 2003, no prelo.

31. Santos, Diana e Paulo Rocha. "AvalON: uma iniciativa de avaliação conjunta para o português", *Actas do XVIII Encontro da Associação Portuguesa de Linguística* (Porto, 2-4 de Outubro de 2002), APL, 2003, no prelo.

32. Santos, Diana. "Against multilinguality", in Stella Neumann & Silvia Hansen-Schirra (eds.), *Proceedings of the workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives* (Lancaster, 27 March 2003), pp. 7-16.

33. Aires, Rachel, Sandra Aluísio, Paulo Quaresma, Diana Santos e Mário Silva. "An initial proposal for cooperative evaluation on information retrieval in Portuguese", 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, Proceedings*, Springer Verlag, 2003, pp. 227-34. [apenas o trabalho das primeira e quarta autoras foi realizado no âmbito do projecto]

34. Mota, Cristina e 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, Proceedings, Springer Verlag, 2003, pp. 184-188.

35. Santos, Diana. "Timber! Issues in treebank building and use", in Nuno J. Mamede, Jorge Baptista, Isabel Trancoso & Maria das Graças Volpe Nunes (eds.), *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, 26-27 June 2003, Proceedings*, Springer Verlag, 2003, pp. 151-158.
36. Santos, Diana, Luís Costa e 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, Proceedings, Springer Verlag*, 2003, pp. 259-266.
37. Almeida, J.J., Alberto Simões, José Castro, Bruno Martins e Paulo Silva. "Projecto TerminUM", in José João Almeida (ed.), *CP3A - Corpora Paralelos, Aplicações e Algoritmos Associados*, Universidade do Minho, 2003, pp. 7-14. [apenas o trabalho dos dois primeiros autores foi realizado no âmbito do projecto]
38. Simões, Alberto. "Alinhamento de corpora paralelos", in José João Almeida (ed.), *CP3A - Corpora Paralelos, Aplicações e Algoritmos Associados*, Universidade do Minho, 2003, pp. 71-77.
39. Sarmento, Luís e Belinda Maia. "Gestor de Corpora - Um ambiente Web integrado para Linguística baseada em Corpora", in José João Almeida (ed.), *CP3A - Corpora Paralelos, Aplicações e Algoritmos Associados*, Universidade do Minho, 2003, pp. 25-30.
40. Maia, Belinda. "The pedagogical and linguistic research implications of the GC to on-line parallel and comparable corpora", in José João Almeida (ed.), *CP3A - Corpora Paralelos, Aplicações e Algoritmos Associados*, Universidade do Minho, 2003, pp. 31-32.
41. Simões, Alberto e J. João Almeida. "NATools -- A Statistical Word Aligner Workbench", Revista da SEPLN - Sociedade Española para el Procesamiento del Lenguaje Natural, a publicar em Setembro de 2003.

De notar que não consideramos aqui os vários artigos escritos sobre recursos da Linguateca, mas cuja redacção não foi financiada pelo presente projecto. Poder-se-iam mencionar, a título de exemplo, sete artigos de Ana Frankenberg-Garcia sobre o COMPARA e três artigos de Eckhard Bick sobre a Floresta Sintá(c)tica. Também não incluímos o já considerável número de artigos que refere trabalho executado sobre o CETEMPúblico ou CETENFolha, por nós criados.

Mencionamos, por outro lado, as apresentações que foram feitas no âmbito da Linguateca, todas elas publicamente acessíveis do nosso portal (referindo-se as [3-16] ao período abrangido pelo projecto).

1. Diana Santos. "Tecnologias da linguagem para o português", apresentação na *Expolíngua Portugal'99*, a 21 de Outubro de 1999.
2. Diana Santos. "Computational Portuguese: the state of the art", apresentação no encontro *Statistical Physics, Pattern Identification and Language Change*, na Universidade de Lisboa, a 14 de Fevereiro de 2000.
3. Diana Santos. Lecture on "The importance of the fagreferent in the information society" at *Fagreferenter 2000*, Biblioteca Universitária de Oslo, a 19 de Junho de 2000.

4. Diana Santos. Lecture on "Evaluation of aligners and alignment", at the Department for British and American Studies at the University of Oslo, 13 September 2000.
5. Diana Santos. "Processamento computacional do português: O que é?". Palestra proferida no Centro de Linguística da Universidade do Porto, a 4 de Outubro de 2000.
6. Diana Santos. "Hands on / Mãos na massa: Fundamentos e dicas para uma linguística experimental". Palestra proferida no Centro de Linguística da Universidade do Porto, a 4 de Outubro de 2000.
7. Diana Santos. Tutorial on "Evaluation of Natural Language Processing systems", at the *Joint International Conference IBERAMIA/SBIA 2000* (Atibaia, São Paulo, Brazil), 19 November 2000.
8. Rachel Aires. "Do you know what you want? An analysis of Web queries in Portuguese", presentation at SINTEF, 20 March 2002.
9. Rachel Aires. "Who am I? What am I doing here?", presentation of PhD work at SINTEF, 24 May 2002.
10. Alexsandro Soares. "Avaliações Conjuntas: Visão Geral", apresentação no *Encontro Preparatório de Avaliação Conjunta do Processamento Computacional do Português* (Universidade do Algarve, 27 de Junho de 2002).
11. Rachel Aires. "Avaliação de Sistemas de Recuperação de Informação (RI): Panorâmica e Reflexões", apresentação no *Encontro Preparatório de Avaliação Conjunta do Processamento Computacional do Português* (Universidade do Algarve, 27 de Junho de 2002).
12. Paulo Rocha e Diana Santos. "Morfoolimpíadas: Uma proposta concreta", apresentação no *Encontro Preparatório de Avaliação Conjunta do Processamento Computacional do Português* (Universidade do Algarve, 27 de Junho de 2002).
13. Diana Santos e Susana Afonso. "Floresta Sintá(c)tica: um recurso para avaliação", apresentação no *Encontro Preparatório de Avaliação Conjunta do Processamento Computacional do Português* (Universidade do Algarve, 27 de Junho de 2002).
14. Diana Santos. "Contrastive Linguistics and NLP", lecture at the Forskerutdanningsseminar om Midler og Maal i Kontrastiv Linvistikk, Arts Faculty, University of Oslo, 26 November 2002.
15. Diana Santos. "The Floresta experience", presentation at the *Swedish Treebank Symposium* (Växjö University, 28-29 November 2002).
16. Diana Santos e Luís Costa. "Linguateca activities", presentation at SINTEF, Oslo, 14 May 2003.
17. Diana Santos. "AvalON'2003: Abertura", apresentação no *AvalON'2003* (Universidade do Algarve, 28 de Junho de 2003).
18. Diana Santos. "Morfolimpíadas: Apresentação detalhada da metodologia e dos problemas identificados", apresentação no *AvalON'2003* (Universidade do Algarve, 28 de Junho de 2003).
19. Cristina Mota. "Avaliação Conjunta de Sistemas de Reconhecimento de Entidades Mencionadas", apresentação no *AvalON'2003* (Universidade do Algarve, 28 de Junho de 2003).

20. Alberto Simões e José João Almeida. "Avaliação de Alinhadores à Frase", apresentação no *AvalON'2003* (Universidade do Algarve, 28 de Junho de 2003).
21. Luís Sarmento e Belinda Maia. "Avaliação de Tradução Automática: Alguns Esforços de Abordagem", apresentação no AvalON'2003 (Universidade do Algarve, 28 de Junho de 2003).
22. Belinda Maia, Anabela Barreiro e Luís Sarmento. "EVAL: Evaluation of Machine Translation at FLUP", apresentação no AvalON'2003 (Universidade do Algarve, 28 de Junho de 2003).
23. Belinda Maia & Luís Sarmento. "Linguateca@Porto", presentation at SINTEF, 10th September 2003.

Finalmente, uma série de outros relatórios e documentação foi produzida na Linguateca, toda ela acessível para consulta, associada a cada projecto desenvolvido. Destaque-se, entre outros, os seguintes:

1. Eckhard Bick, Susana Afonso e Ana Raquel Marchi. s/d "Notational and terminological guide-lines", http://www.visl.hum.sdu.dk/visl/pt/guidelines.html.
2. Eckhard Bick, Susana Afonso e Ana Raquel Marchi. s/d "Documentation of the choices in the treebank project", http://visl.hum.sdu.dk/visl/pt/MainDocumentationTreebank.html.
3. Afonso, Susana e Ana Raquel Marchi. "Critérios de separação de sentenças/frases". Floresta Sintá(c)tica, 28 de Fevereiro de 2001, http://acdc.linguateca.pt/treebank/CriteriosSeparacao.html.
4. Afonso, Susana e Ana Raquel Marchi. "A etiqueta <sic> </sic>". Floresta Sintá(c)tica, 28 de Fevereiro de 2001, http://acdc.linguateca.pt/treebank/CriteriosSic.html.
5. Afonso, Susana Cavadas. "Na trilha de um Teste Inter-Anotadores". Floresta Sintá(c)tica, 8 de Novembro 2001. http://acdc.linguateca.pt/treebank/TrilhaTIA.ps.
6. Haber, Renato Ribeiro. "Pica-pau: Um protótipo de ferramenta para visualização e edição de árvores sintáticas". Floresta Sintá(c)tica, 6 de Novembro de 2001, http://acdc.linguateca.pt/treebank/Picapau.html.
7. Rocha, Paulo. "Gestão das páginas do projecto Processamento Computacional do Português", 19 de Novembro de 2001.
8. Afonso, Susana. "Manual do anotador da Floresta Sintá(c)tica", Maio de 2002.
9. Frankenberg-Garcia, Ana. "COMPARA – Paragraph alignment instructions", 19 de Setembro de 2002.
10. Frankenberg-Garcia, Ana. "COMPARA – instruções de alinhamento por parágrafo", 25 de Setembro de 2002.
11. Frankenberg-Garcia, Ana e Diana Santos. "COMPARA – editar o alinhamento por frase", 25 de Setembro de 2002.
12. Frankenberg-Garcia, Ana e Diana Santos. "COMPARA – editing sentence alignment", 25 de Setembro de 2002.
13. Frankenberg-Garcia, Ana. "COMPARA – Optical Character Recognition (OCR) editing and preliminary markup instructions", 25 de Setembro de 2002.
14. Frankenberg-Garcia, Ana. "COMPARA – correcção do reconhecimento óptico de caracteres (roc) e etiquetagem preliminar", 25 de Setembro de 2002.

15. Costa, Luís, Cristina Mota e Luís Sarmento. "Alguns comentários à usabilidade do serviço COMPARA", 4 de Dezembro de 2002.

16. Santos, Diana. "Atomização e separação de frases". Projecto AC/DC, 16 de Janeiro de 2003.

17. Santos, Diana. "Anotação dos corpora". Projecto AC/DC, 13 de Fevereiro de 2003.

18. Aires, Rachel Virgínia Xavier. "Linguarudo - Uma Arquitetura Lingüisticamente motivada para Recuperação de Informação de textos em português", Monografia apresentada ao Instituto de Ciências Matemáticas de São Carlos - USP, para o Exame de Qualificação, como parte dos requisitos para a obtenção do título de Doutor em Ciências - Área de Ciências de Computacção e Matemática Computacional, Março de 2003.

Índice

Apresentação....1
O que é a Linguateca?....1
Pressupostos básicos da Linguateca....2
Público-alvo....2
Principais actividades....2
Estrutura do relatório....3
Eixo 1: Disseminação e informação....5
Algumas medidas do impacto e da grandeza do portal....6
Busca: Sistema de procura no portal....8
Menuseador: Sistema de gestão do portal....9
Eixo 2: Disponibilização e criação de recursos....11
Projecto AC/DC, www.linguateca.pt/ACDC/....12
Projectos CETEMPúblico e CETENFolha....16
www.linguateca.pt/cetempublico/....16
www.linguateca.pt/cetenfolha/....16
Projecto COMPARA/DISPARA, www.linguateca.pt/COMPARA/....18
Projecto Floresta Sintá(c)tica, www.linguateca.pt/Floresta/....20
AnELL....21
Corpógrafo: Gestor de corpora para ensino de tradução e terminologia....22
Serviços relacionados com a avaliação de tradução automática....23
METRA: MEtaTRAdutor automático....23
Boomerang: Determinação de Pontos de Fixação em Motores de Tradução....23
TrAva: Compilação de julgamentos de qualidade sobre traduções automáticas....24
DISPARA....25
Águia....26
Corpora criados de raiz....27
Programas de observação do uso dos serviços....28
Repositório....28
Mini-serviços....28
Programas....29
Eixo 3: Avaliação conjunta....31
Descrição da comunidade associada à onda de avaliação conjunta....32
Morfolimpíadas....32
Avaliação de reconhecimento de entidades mencionadas....33
Avaliação de recolha de informação ("information retrieval")....34
Avaliação de tradução automática....34
Avaliação de alinhamento....34
Avaliação de análise sintáctica....34
Outras actividades....34
Actividades de investigação e formação....35
Equipa associada à Linguateca....37
Publicações e palestras....39
Índice....46