Breve panorâmica dos recursos de português mencionados na Web

Signe Oksefjell e Diana Santos
projecto@informatics.sintef.no
http://www.portugues.mct.pt/
Processamento computacional do português,
SINTEF Telecom and Informatics,
Boks 124 Blindern, N-0314 Oslo, Noruega

Este documento pretende apresentar de forma resumida a informação presente na nossa página de recursos de português computacional na Web, **http://www.portugues.mct.pt/recursos.html**, mostrando de uma forma mais estruturada algumas informações pertinentes e que serão pormenorizadas, para os recursos disponíveis, numa futura página de avaliação, que planeamos também tornar acessível na Web.

Contexto

O trabalho que apresentamos aqui faz parte de uma iniciativa mais vasta do Ministério da Ciência e da Tecnologia português com vista à dinamização e progresso do processamento computacional da língua portuguesa. Uma das primeiras actividades desse programa de acção é catalogar o que já existe e traçar o perfil da área em Portugal e no mundo. Por seu lado, uma das componentes desta actividade é identificar a presença do processamento do português na rede, criando um catálogo que poderá beneficiar todos quantos queiram desenvolver sistemas que processem a nossa língua. A presente versão desse catálogo, contendo mais de 200 ponteiros, acessível do nosso servidor http://www.portugues.mct.pt/, é o ponto de partida para o presente artigo, que tenta apresentar o seu conteúdo e tecer algumas considerações sobre a sua criação.

Convém indicar que esta panorâmica foi feita **sem consulta directa aos actores e responsáveis pelos projectos**, e só se refere a sistemas **mencionados na WWW**. Muitos dos projectos e sistemas existentes em Portugal, no Brasil ou noutros países – e de que temos conhecimento – não se encontram portanto mencionados. Tentámos contudo referir, na própria página, listas e catálogos que contivessem recursos de outra índole, tais como Castilho et al. (1995), Correia (1995), Martins et al. (1998) e Nascimento et al. (1995). Por outro lado, seria loucura tentar reclamar qualquer exaustividade, embora esse seja o nosso objectivo teórico: todos os dias descobrimos outros caminhos, provavelmente também todos os dias nasce uma página nova na Web relevante para o processamento da língua portuguesa…

Apesar da consulta directa ter faltado, tentámos publicitar, a nível do processamento do português (lista do Forum-LP) e de listas internacionais (Corpora, Linguist, etc.) a nossa iniciativa, pedindo contribuições, sugestões e críticas – fizemos o primeiro anúncio em fins de Julho, e anunciámos a renovação das páginas em princípio de Setembro do corrente ano (1998). É talvez interessante referir, como dado sociológico, que tivemos muito pouca resposta do lado da comunidade que pretendíamos servir. Apesar de catalogarmos mais de 200 páginas, tivemos até agora apenas 18 respostas de parabéns e encorajamento ou com sugestões e adições, ainda que tivéssemos tido, desde a criação da primeira versão, mais de 1200 visitas externas, pela Web, provenientes de mais de 400 computadores diferentes.

Resenha da informação presente em recursos.html

Em seguida apresentamos um resumo do conteúdo da página, dividida em quatro secções – corpora, dicionários e bases de dados terminológicas, ferramentas e outros recursos – tentanto concentrar a informação pertinente em forma de tabelas. Um asterisco (*) indica que a informação correspondente não se encontra acessível. Para identificar devidamente os recursos a que nos referimos apenas pelo

nome, o leitor tem de usar a própria página da Web, que resumimos aqui.

Corpora

Na lista de corpora, representada na tabela 1 – a única tabela cuja informação foi em alguns casos complementada com a informação presente em Nascimento et al. (1995) –, encontram-se ponteiros para 16 corpora que ou são monolingues para o português (8) ou bilingues/multilingues que incluem o português (8). 10 são de português escrito (o CRPC incluindo uma parte de português falado, transcrito), 5 de fala (o CORAL também é transcrito) e 3 de português falado, e transcrito.

Os únicos corpora disponíveis directamente na WWW são o Corpus jornalístico Natura-PUBLICO, desenvolvido na Universidade do Minho com a autorização do jornal Público, e os corpora desenvolvidos no âmbito do projecto CHILDES, acessíveis dos servidores internacionais desse projecto. O Wordtheque permite a procura de palavras na rede, estando depois cada texto em que a palavra ocorre integralmente acessível. Quatro dos corpora multilingues (EUROM 1, ECI/MCI, MLCC e European Language Newspaper Text – este último apenas disponível para membros do LDC em 1995 e 1996) são distribuídos na forma de CD-ROM. Os restantes nove (9) corpora não estão disponíveis para consulta fora das instituições responsáveis.

Tabela 1

Tipo	**Nome do corpus**	**Tamanho (só português)**	**Tamanho previsto/data**
escrito			
	Córpora de Português Medieval	620.000 palavras (Maio 1996)	* / 1999
	CORPUS		40 milhões de palavras / 1997
	Corpus de Referência do Português Contemporâneo (CRPC)	40 milhões de palavras	
	Corpus jornalístico Natura-PUBLICO	1,76 Gb	
	ECI/MCI	675.000 palavras	
	English-Norwegian Parallel Corpus	180.000 palavras	
	European Language Newspaper Text	15 milhões de palavras	
	MLCC	8 milhões de palavras	
	Tycho Brahe Parsed Corpus of Historical Portuguese		1 milhão de palavras / 1999
	Wordtheque	266 (extractos de) obras	
fala			
	BD-FALA	2 Gb	
	BD-PUBLICO	10 milhões (2 Gb)	
	CORAL		1,8 Gb / 1998
	EUROM 1	3 Gb (parte portuguesa)	
	SPEECHDAT	*	
falado			
	CHILDES (Batoréo, etc.)	120 narrativas (pelo menos)	513kB (pelo menos)
	CORAL		1,8Gb / 1998
	CRPC	1,5milhões de palavras	

Dicionários e léxicos

Nesta lista encontram-se ponteiros para 94 dicionários ou outros recursos lexicais. Existe uma grande variedade de recursos nesta lista, principalmente quanto a tamanho e tema. A tabela 2 fornece uma primeira sistematização, que é depois pormenorizada para cada categoria nas tabelas seguintes, referentes aos 16 dicionários monolingues gerais, 46 dicionários multilingues ou bilingues, e 32 dicionários monolingues especializados e bases de dados terminológicas. PE e PB significam respectivamente português europeu e brasileiro.

Dez (10) dos dicionários são multimédia: Aurelinho – Dicionário Infantil Multimídia, Diciopédia, Dicionário multilingue interactivo, Dicionário multimédia universal (por-fra/fra-por e por-ing/ing-por), Dicionário Visual Verbo, DIC Michaelis multimídia e Grande

Dicionário Multimédia Universal de Língua Portuguesa.

Em relação à categoria de dicionários multilingues e bilingues, convém salientar que a maior parte são bilingues. Apenas cinco destes dicionários são apresentados como multilingues: Ergane, The Internet Dictionary Project, Dicionário Eletrônico ZAZ!, Dicionário LOGOS, Dicionário multilingue interactivo.

Por outro lado, na categoria dos dicionários monolingues especializados e bases de dados terminológicas, a maior parte destas colecções são monolingues. Há, contudo, dez (10) bi- ou multilingues: ACRONYMS, Dicionário de termos artísticos, DIC MICHAELIS Técnico de termos técnicos, DIC MICHAELIS Técnico Direito & Economia, EURODICAUTOM, EUTERPE, Fish terminology, Glossary of Portuguese Narcotics Terms, LORETO e VERBA, e duas com as entradas em português e a explicação em inglês: Abbreviations e Glossary of Portuguese Narcotics Terms.

Na tabela 5 indicamos a negrito os (poucos) casos em que a especialização do dicionário não se refere à área, mas sim à função ou tipo de linguagem. Quando o tamanho aparece sublinhado, significa que foi contado por nós.

Tabela 2

Tipo de dicionário	**Quant.**	**Totalmente acessível**	**Acessível para consulta**		**Venda**	**Não acessível**
			livre	paga		
dicionários gerais	16	3	1	3	7	2
dicionários especializados e BDs terminológicas	32	20	4		8	
dicionários multi- ou bilingues	46	2	18	10	16	

Dicionários monolingues gerais (16)

Freeware:		Dicionário br.ispell (PB), Dicionário português (Ispell) (PE), Dicionário português (Jspell) (PE)
Consulta	livre:	Dicionário da Língua Portuguesa On-Line (PE)
	paga:	Dicionário Michaelis Soft Executivo, Melhoramentos Soft da Língua portuguesa, Nova Enciclopédia Ilustrada Folha (PB)
Venda:	CD:	AURÉLIO Eletrônico (PB), Dicionário da Língua Portuguesa On-Line (PE), Dicionário PROfissional da Língua Portuguesa (PE), Diciopédia (PE), Dicionário Visual Verbo (PE), DIC MAXI Michaelis português (PB), Grande Dicionário Multimédia Universal de Língua Portuguesa (PE)
	disquete:	Dicionário Básico da Língua Portuguesa (PE)
Não acessível:		Monolingual Portuguese lexicon (PE), Portuguese morphological lexicon Palavroso (PE)

Tabela 3

Dicionários monolingues gerais	**Tamanho (verbetes)**	**Origem**
AURÉLIO Eletrônico	130.000	Lexikon Informática
Dicionário Básico da Língua Portuguesa	17.000	Porto Editora / Priberam
Dicionário br.ispell	*	Ricardo Ueda Karpischek, Univ. de S. Paulo[1]
Dicionário da Língua Portuguesa On-Line	500.000	Porto Editora / Priberam
Dicionário Michaelis Soft Executivo	*	Michaelis
Dicionário português (Ispell)	40.000	Projecto Natura, Univ. Minho
Dicionário português (Jspell)	45.000	Projecto Natura, Univ. Minho
Dicionário PROfissional da Língua Portuguesa	500.000	Porto Editora / Priberam
Diciopédia	90.000	Porto Editora /

[1] O projecto foi desenvolvido a nível individual, sem apoio institucional da Universidade, que no entanto fornece a infraestrutura informática e de rede.

		Priberam
Dicionário Visual Verbo	*	Editorial Verbo
DIC MAXI Michaelis Português	200.000	Michaelis
Grande Dicionário Multimédia Universal de Língua Port.	500.000	Texto Editora
Melhoramentos Soft da Língua portuguesa	*	Melhoramentos
Monolingual Portuguese lexicon	60.000	Centro de Linguística Univ. Lisboa
Nova Enciclopédia Ilustrada Folha	*	Empresa Folha da Manhã S/A
Portuguese morphological lexicon Palavroso	60.000	INESC

Dicionários multilingues e bilingues (46)

Totalmente acessíveis: Ergane, The Internet Dictionary Project

Consulta livre: Dicionário Eletrônico ZAZ! (PB), Dicionário LOGOS, Travlang's Translating Dictionaries para 8 línguas (16)

paga: Dicionários Michaelis dicionários bilingues para 5 línguas (10)

Venda: CD: Dicionário de Caboverdiano-Português, Dicionário multilingue interactivo, Dicionário Multimédia Universal Português/ Francês, Francês/ Português, Dicionário Multimédia Universal de Português-Inglês, Inglês-Português, Dicionário PROfissional de inglês-português, Dicionário PROfissional de português-inglês[#], DIC MAXI Michaelis multimídia (6 idiomas), DIC Michaelis (2 idiomas), Online Portuguese-English/English Portuguese dictionary, Webster's inglês-português, português-inglês

disquete: Dicionário PROfissional de inglês-português, Dicionário PROfissional de português-inglês[#]

download: Portuguese-English Dictionary Macintosh, Portuguese-English Dictionary – Windows

[#]Estes recursos encontram-se em ambos os formatos: CD e disquete

Tabela 4

Dicionários multilingues/bilingues	**Tamanho (verbetes)**	**Origem**
Dicionário de Caboverdiano-Português	4.000	Priberam / Verbalis
Dicionário eletrônico ZAZ!	*	Livraria Nobel
Dicionário LOGOS	7.580.560 verbetes (todas as línguas)	LOGOS
Dicionário multilingue interactivo	10.000 (cada língua)	Porto Editora
Dicionário multimédia universal (por-fra/fra-por)	20.000 (cada língua)	Texto Editora
Dicionário multimédia universal (por-ing/ing-por)	20.000 (cada língua)	Texto Editora
Dicionário PROfissional (por-ing/ing-por)	82.000 / 68.000	Porto Editora / Priberam
Dicionários Michaelis (para 5 línguas)	*	Michaelis
DIC MAXI Michaelis multimídia (6 idiomas)	319.000 verbetes	Michaelis
DIC Michaelis (2 idiomas)	*	Michaelis
Ergane	14.547	Travlang
Internet dictionary project	*	June29
Online Portuguese-English/English Portuguese dict.	1.500 / 1.900	Online Dictionaries
Portuguese-English dictionary Macintosh/Windows	40.000 (cada língua)	Exceller Software Corp
Travlang's translating dictionaries (para 8 línguas)	entre 2.500 e 14.500 (cada língua)	Travlang
Webster's inglês-português/português-inglês	40.000	Webster's

Dicionários monolingues especializados e bases de dados terminológicas (32)

Totalmente acessíveis: Abbreviations and acronyms used in the Portuguese-language press, ACRONYMS / SIGLAS / ABREVIATURAS, The Alternative Portuguese Dictionary (Portuguese slang) (PE, PB), Dicionário Alagoano (PB), Dicionário de Aquarismo, Dicionário de Astronomia e Áreas Afins (PB), Dicionário do internetês, Dicionário informática, Dicionário interativo – informática e internet (PB), Dicionário interativo de química;

Internet e Multimídia (PB), Fish terminology, GLOSSÁRIO – A Reforma da União europeia, Glossário da gíria da internet (PB), Glossário de Áudio e Vídeo, Glossário de informática, Glossary of Portuguese Narcotics Terms, List and Glossary of medical terms, Lista de provérbios (PE), Microsoft glossaries – Brpor e Eupor

Consulta: livre: Dicionário de calão (PE), Dicionário de termos artísticos (PB), EURODICAUTOM, EUTERPE, LORETO, VERBA

Venda: CD: Aurelinho – Dicionário Infantil Multimídia (PB), Dicionário Verbo de Inglês Técnico e Multimédia (PE), DIC MICHAELIS Técnico de termos técnicos (PB), DIC MICHAELIS Técnico Direito & Economia (PB), Speri-Data AG Basic dictionaries (colloquial language) (PE)

disquete: Dicionário de sinónimos do FliP (PE), LORETO e VERBA

Tabela 5

Dicionários ou BDs especializados	**Tamanho (termos)**	**Área**
Abbreviations and acronyms used in the Portuguese-language press	1.918	**abreviaturas e acrónimos** (português-inglês)
ACRONYMS / SIGLAS / ABREVIATURAS	a contar	**abreviaturas e acrónimos** (inglês e português)
Alternative Portuguese Dictionary	51	calão
Aurelinho	*	**linguagem infantil**
Dicionário Alagoano	864	**linguagem regional**
Dicionário de Astronomia e Áreas Afins	58	astronomia
Dicionário de calão	*	calão
Dicionário de sinónimos do FliP	*	**sinónimos**
Dicionário de termos artísticos	3.300 (acessíveis ~100)	arte
Dicionário do internetês	208	Internet
Dicionário informática; Internet e Multimídia	247	informática
Dicionário interativo de química	673	química
Dicionário Verbo de Inglês Técnico e Multimédia	120.000	informática, física nuclear, genética molecular, botânica, etc.
DIC MICHAELIS Técnico de termos técnicos	20.000	termos técnicos e científicos (6 idiomas)
DIC MICHAELIS Técnico Direito & Economia	64.000	direito e economia (português-alemão-português)
EURODICAUTOM	339.362	terminologia e abreviaturas da Comissão da UE
EUTERPE	150.000	Terminologia do Parlamento Europeu
Fish terminology	29	nomes de peixes em 9 línguas
GLOSSÁRIO – A Reforma da União europeia	150	temas da reforma da UE
Glossário de Áudio e Vídeo	691	termos técnicos de áudio e vídeo
Glossário de informática	329	informática
Glossary of Portuguese Narcotics Terms	1.288	droga (português-inglês)
List and Glossary of medical terms	3.651	termos médicos
Lista de provérbios	500	provérbios
LORETO	800	biotecnologia, meio-ambiente, energia, telecomunicação, etc.
Microsoft glossaries (Brpor e Eupor)	34.270 (PB) 50.463 (PE)	terminologia Microsoft
Speri-Data AG Basic dictionaries	9.000	**linguagem coloquial**
VERBA	83.000 (as 6 línguas)	politecnologia

Ferramentas computacionais (38)

As ferramentas estão divididas em 5 categorias:

- Ajuda à redação (13 ponteiros)
- Componentes básicos de um sistema de PLN: analisadores ou geradores da língua (12 ponteiros)
- Tradução automática (9 ponteiros)
- Síntese de fala (2 ponteiros)
- Ajuda ao ensino (2 ponteiros)

Na primeira categoria (**Ajuda à redacção**), resumida na tabela 6, encontram-se principalmente correctores ortográficos, mas também correctores gramaticais em geral. Os correctores ortográficos ISPELL (PE), br.spell (PB) e Domínio (PB) encontram-se como freeware na WWW. As ferramentas para a língua portuguesa (FliP), que incluem um corrector grammatical, um corrector ortográfico, um editor de texto e um hifenizador, são vendidas em forma de disquete. A Gramática Eletrônica (PB), o LEXIKON (PB), as duas versões de Orthográphos (PB), o Revisor gramatical DTS (PB) e a Redação Língua Portuguesa (PB) são distribuídos na forma de CD. Ainda na lista Lince (PE) – já não existente no mercado – e ReGra (PB) – integrado num editor de texto comercial.

Não estão incluídos nesta lista a maior parte dos correctores ortográficos incluídos em sistemas proprietários (como por exemplo editores de texto), por não serem geralmente descritos independentemente (e não se encontrarem por isso facilmente na rede).

Na categoria **Componentes básicos de um sistema de Processamento de Linguagem Natural: analisadores ou geradores da língua**, descrita na tabela 7, há mais variedade em tipo de ferramenta, mas em primeiro lugar encontram-se aqui analisadores ou geradores morfológicos (6) e analisadores sintácticos (4) da língua portuguesa, um alinhador inglês-português, e um sistema para ajudar a análise manual ("tagging") de corpora. Três destes são freeware e cinco podem experimentar-se via Internet. Finalmente, inluímos aqui também um programa para desenvolver dicionários (Ergane).

Na tabela 8, há 9 ponteiros para programas de **Tradução automática**. 6 são multilingues e incluem o português como língua de partida e de chegada, enquanto 3 são bilingues, e dirigidos para o par em causa: português-dinamarquês (Portdan) e inglês-português (DIC TRADUTOR e GEVER). Quatro são acessíveis através da Internet para textos pequenos. GEVER é freeware, enquanto o Universal Translator de Luxe pode ser experimentado durante trinta dias sem pagar. Finalmente, o Word translator, como o nome indica, é um tradutor palavra-a-palavra.

Outra categoria, na tabela 9, é a de **Síntese de fala**, onde se encontram dois sintetizadores: o DIXI, que é um sintetizador de fala a partir de texto em português, e o SVITD, que é um sintetizador de números de telefone em português. Os dois encontram-se disponíveis para teste na WWW; o segundo é, além disso, usado nos serviços da Portugal Telecom (número de telefone 118).

Na última categoria, **Ajuda ao ensino**, há 2 ponteiros, dispostos na tabela 10: para a Interactive grammar, disponível para consulta e para a Verboteca portuguesa, disponível no seu conjunto para download. De notar que só se encontram aqui mencionadas ferramentas cujo objectivo seja exactamente "Computer-aided learning", outros programas e recursos também poderiam evidentemente aqui ser integrados como auxiliares nesta tarefa.

Tabela 6: ajuda à redacção

Como obter	Ferramenta	Desenvolvido por
Venda		
	Domínio	Alania Laboratório Digital
	FliP	Priberam / Porto Editora
	Gramática eletrônica	Lexikon Informática
	LEXIKON	Lexikon Informática
	Orthográphos	DTS Software
	Orthográphos Maxi (3 idiomas)	DTS Software
	Redação Língua Portuguesa	*
	ReGra (parte de REDATOR)	Univ. de São Paulo
	Revisor gramatical DTS	DTS Software
Download		
	br.ispell	Ricardo Ueda Karpischek, Univ. de São Paulo[2]
	Domínio	Alania Laboratório Digital
	ISPELL	Projecto Natura, Univ. do Minho
Não acessível		
	Lince	ILTEC / Priberam

Tabela 7: analisadores e geradores de português

Disponibilidade	Ferramenta	Desenvolvido por
Consulta www		
	Automatic Analysis of Portuguese	Eckhard Bick, Univ. de Aarhus
	ConVer – Conjugação Verbal	Univ. Federal do Rio Grande do Sul, Instituto de informática
	Finite-State Morphological Analyzer	Xerox research
	LOGOS Universal conjugator	LOGOS
	XRCE Part of Speech Disambiguators	Xerox research
Download		
	Conjugador de verbos da língua portuguesa	Ricardo Ueda Karpischek, Univ. de São Paulo[2]
	ConVer – Conjugação Verbal	Univ. Federal do Rio Grande do Sul, Instituto de informática
	Ergane	Gerard van Wilgen, Travlang
	Etiquetador para português	Projecto Natura, Univ. do Minho
	JSPELL: analisador morfológico	Projecto Natura, Univ. do Minho
	LS-GRAM Grammars for EU Languages	União Europeia / ILTEC
	Tagging Aid Tool of the Tycho Brahe Corpus	C. de Menezes, Univ. de São Paulo
	The Translation Corpus Aligner	Knut Hofland, Univ. de Bergen

Tabela 8: tradutores automáticos

Disponibilidade	Ferramenta	Línguas	Desenvolvido por
Venda			
	Comprende	inglês-português-inglês	Globalink
	DIC TRADUTOR	inglês-português	DTS Software
	SYSTRAN®Personal	inglês-português-inglês	SYSTRAN
	Universal Translator de Luxe	25 línguas[2]	Language Force
	Word translator	20 línguas[3]	Translation Experts
Consulta www			
	Altavista Translation Service	inglês-português-inglês	SYSTRAN
	Portdan – trans. into Danish	português-dinamarquês	Eckhard Bick, Univ. de Aarhus
Shareware			
	E-mail Translator Plug-In for Eudora	inglês-português-inglês	Globalink
Freeware			
	GEVER	inglês-português	Vilson Leffa, Univ. Católica de Pelotas

Tabela 9: sintetizadores de fala

Ferramenta	Desenvolvido por
DIXI	INESC / CLUL
SVIDT	INESC / Portugal Telecom

Tabela 10: ajuda ao ensino

Ferramenta	Desenvolvido por
Interactive grammar	Eckhard Bick, Univ. de Aarhus
Verboteca portuguesa	Fernando Moura, Universidade Católica de Lovaina

Outros

Não tratámos aqui a categoria "Outros recursos", presente na nossa página da rede, por não nos parecer pertinente uma maior estruturação. Contentamo-nos com indicar que tem ponteiros para:

- Catálogos de literatura em português acessíveis na WWW (2)
- Textos em português (27)
- Material didáctico (8)

[2] Chinese, Czech, Danish, Esperanto, French, German, Greek, Hungarian, Italian, Indonesian, Latin, Japanese, Korean, Norwegian, Portuguese, Romanian, Russian, Slovak, Spanish, Swahili, Swedish, Thai, Turkish, Ukrainian, and Vietnamese

[3] Português, inglês, alemão, francês, espanhol, italiano, grego, finlandês, norueguês, sueco, dinamarquês, japonês, húngaro, checo, polaco, holandês, islandês, croata, sérbio, romeno

- Informação sobre a língua portuguesa (5)
- Outros catálogos com links relacionados com a língua portuguesa (5)
- Iniciativas políticas, relacionadas com o português na era da informação (3)

Outros parâmetros de categorização

Enquanto que é frequente que os dicionários (e por vezes os corpora) mencionem a variante (PE, PB ou português medieval) a que se referem, tal é no caso das ferramentas computacionais raramente explicitado, o mesmo se passando na categoria "Outros recursos".

Tentámos por isso identificar, para o caso das ferramentas, a distribuição entre as duas variantes mais tratadas, ou seja, o português de Portugal e o do Brasil, identificados pelos adjectivos "europeu" e "brasileiro" nas tabelas 11 a 13.

Da mesma forma, tentámos classificar os corpora de português escrito em relação precisamente à variedade de português (norma) que reflectem, na tabela 14, cujas iniciais correspondem directamente aos corpora já citados na tabela 1. Os corpora de fala e de português falado que conseguimos encontrar na rede eram exclusivamente de português europeu. "PA" indica "português antigo".

Tabela 11

Ferramentas de ajuda à redacção	
Português europeu	**Português brasileiro**
FliP	Domínio
Ispell	Gramática eletrônica
Lince	Lexikon
	Orthográphos
	Orthográphos Maxi
	ReGra
	Revisor gramatical DTS
	br.ispell

Tabela 12

Analisadores ou geradores da língua		
Português europeu	**Português brasileiro**	**Português europeu/brasileiro**
Etiquetador para português	Automatic Analysis of Portuguese	Finite-State Morphological Analyzer
Jspell	ConVer	LOGOS Universal conjugator
LS-Gram	Conjugador de verbos da língua portuguesa	XRCE Part of Speech Disambiguators
		Ergane
		Tagging Aid Tool of the Tycho Brahe Corpus
		The Translation Corpus Aligner

Tabela 13

	Ferramentas de tradução automática	
Português europeu	**Português brasileiro**	**Difícil de determinar**
	DIC Tradutor	Comprende
	Portdan	SYSTRAN
	GEVER	Universal translator de luxe
		Word translator
		Altavista translation service
		E-mail translator plug-in for Eudora

Tabela 14

Corpus	**PE**	**PB**	**PA**
CPM			100%
CORPUS	100%		
CRPC	94%	5%	
CJ Natura-PUBLICO	100%		
ECI/MCI		100%	
ENPC	62,5%	37,5%	

ELNT	100% (estimamos)		
MLCC	100%		
TBPCHP			100%
Wordtheque	51%	49%	

Observações finais

Este texto não pretende ser mais do que uma primeira "fotografia", precisamente datada (de 7 de Outubro de 1998), da informação sobre o processamento computacional da língua portuguesa na Web. Ao contrário de fixar uma área, espera-se que motive os seus investigadores a produzirem e a colocarem mais informação na rede, ou seja, que dê origem a uma explosão de sítios com recursos e informação sobre o processamento computacional da nossa língua.

Pensamos que o catálogo a que nos temos dedicado servirá para identificar o trabalho já feito, assim como as áreas mais necessitadas de investigação e de trabalho prático, além de permitir uma reutilização efectiva dos recursos existentes e facilitar uma comunicação alargada entre os membros da comunidade científica.

Por outro lado, as limitações do presente texto, e das páginas da rede a que se referem, demonstram claramente que não é trivial usar uma metalinguagem comum, e que as classificações, a nomenclatura, e a importância relativa da informação que as páginas para as quais apontamos expõem varia de forma extrema. Algumas questões são contudo de salientar:

- em muitos casos, não concordamos com a terminologia usada pelos donos dos recursos, mas cingimo-nos a ela
- as listas de recursos são sempre apresentadas por ordem alfabética; no presente artigo, o mesmo recurso aparece por vezes em várias posições numa mesma tabela
- em paralelo, desenvolvemos um catálogo de actores (por agora restrito a projectos e grupos) na área do processamento computacional da língua portuguesa, cujo endereço é http://www.portugues.mct.pt/actores.html; nessa página encontram-se projectos de desenvolvimento de ferramentas (ou dicionários, ou corpora) que não são mencionados no catálogo de recursos por parecer não existir ainda (ou já não) o recurso a que se referem
- não é porventura possível avaliar, com base na descrição das páginas, a classificação que os seus autores consideram mais justa, por isso é preciso contar com a colaboração da comunidade para melhorar o nosso catálogo; por outro lado, resta-nos a esperança de que a nossa tentativa de classificação possa ajudar os donos de páginas que se sentiram injustiçados a reformular de certa forma o seu conteúdo, de forma a tornar mais óbvio o que têm para oferecer e que foi por nós mal compreendido

Ao tentarmos uma apresentação quase linear, na página da Web, esperamos que o nosso trabalho possa servir de ponto de partida para outras classificações, dando origem a uma pluralidade de emaranhados cuja estrutura seja mais clara e mais útil para necessidades distintas. Graças ao carácter eminentemente democrático da Web, nunca o nosso catálogo poderá ser mais do que **uma** possibilidade de apresentar informação que nos parece relevante, e que já existe à espera de ser encontrada.

No presente texto, além de apresentar alguns dados não directamente acessíveis na página que estamos a publicitar, tais como o número de entradas dos dicionários ou o tamanho dos corpora, tentámos esboçar, de uma maneira incipiente, outras formas simples de estruturar essa mesma informação, como por exemplo a classificação segundo a variante do português. Outras critérios óbvios de classificação seriam a plataforma informática e o estatuto legal ou de disponibilidade, este último indicado na página através de um código de imagens.

Ainda que a forma em papel deste artigo cedo ficará desactualizada, tencionamos ir melhorando a informação presente no nosso servidor, quer em profundidade, à medida que formos recebendo e observando os produtos já encomendados, quer em extensão, renovando aqui o apelo a todos os produtores e investigadores na área. Tal apelo é feito muito especialmente a todos os investigadores do

Brasil, que certamente ficou menos coberto nesta primeira versão do catálogo, devido à origem das autoras e do seu financiamento. Esperamos sinceramente que tal possa ser corrigido com a boa-vontade dos investigadores brasileiros em nos apontar as suas páginas e projectos, como já aconteceu em muitos casos que aqui publicamente agradecemos.

Referências

Castilho, Ataliba Teixeira de, Giselle Machline de Oliveira e Silva e Dante Lucchesi. "Informatização de acervos da língua portuguesa" em Nascimento, Maria Fernanda Bacelar do, Maria Celeste Rodrigues e José Bettencourt Gonçalves (orgs.), *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística, Vol. 1 Corpora*, Lisboa, 1995, pp.113-128.

Correia, Margarita "Dicionários de Língua Portuguesa: Lista não Exaustiva de Títulos Disponíveis", em Faria, Isabel Hub e Margarita Correia (orgs.), *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística, Vol. II Dicionários*. Lisboa, 1995, pp.279-286.

Martins, Ciro, Isabel Mascarenhas, Hugo Meinedo, João Neto, Luís Oliveira, Carlos Ribeiro, Isabel Trancoso e Céu Viana. "Spoken Language Corpora for Speech Recognition and Synthesis in European Portuguese", http://www.speech.inesc.pt/bib/Trancoso98a/poster.html

Nascimento, Maria Fernanda Bacelar do, Maria Celeste Rodrigues e José Bettencourt Gonçalves (Orgs.). "Corpora portugueses", *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística, Vol. 1 Corpora*, Lisboa, 1995, pp.423-447.