Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva

Diana Santos
Linguateca, SINTEF ICT
Diana.Santos@sintef.no

Resumo

Este artigo faz um balanço pessoal do percurso da Linguateca, uma organização virtual em demanda de uma maior facilidade e qualidade no processamento da língua portuguesa, nos últimos dez anos.

Inicio o artigo por uma curta perspectiva histórica para explicar o contexto em que a Linguateca surgiu e quais os objectivos iniciais para o progresso da área. Avalio de seguida resumidamente a situação actual no que respeita a esses objectivos iniciais, bastante vagos, identificando o que foi cumprido e perspectivando o que ficou por fazer.

Aproveito também a oportunidade para apresentar as variadas inflexões que o projecto tomou, num percurso que não foi linear.

Faço depois uma breve excursão pelos principais pontos atingidos, mas sem a preocupação de ser exaustiva, dado que o texto não pretende ser um relatório, mas sim uma reflexão crítica sobre o processo e os resultados, tentando relacioná-la, sempre que possível, com a discussão pública que teve lugar dez anos volvidos no Encontro *Linguateca: 10 anos*, em Aveiro a 11 de Setembro de 2008.

Embora o artigo seja centrado sobre a Linguateca, tento fazer numa última secção algumas pontes com outro trabalho em processamento do português, de forma a não transmitir a ideia errada de que teríamos sido os únicos a trabalhar na área ou a progredir neste período.

Termino o artigo com uma breve secção com algumas sugestões para projectos que possam continuar o espírito da Linguateca ou reforçar as contribuições da Linguateca para o objectivo mais geral da dignificação e da melhoria do processamento da língua portuguesa.

O processo de tornar o processamento do português mais percorrido e mais agradável assemelha-se ao desbravamento de vários caminhos num emaranhado de questões e problemas semelhante a uma selva ou país – daí o título deste texto referir o "mapa da portuguesificação". Ao invés de considerar o trabalho concluído, ponho a tónica no muito ainda que é preciso fazer nesta área, em que a acção da Linguateca é (ou foi) comparável, apenas, à criação de alguns caminhos. Também por isso indico neste texto aquelas sendas que acabaram em becos sem saída, mas que aumentaram a nossa experiência ou nos convenceram de que não devíamos seguir por ali.

1 Apresentação

A Linguateca foi um projecto político-científico financiado pelas autoridades na área da ciência e da tecnologia em Portugal para tratar do processamento computacional da língua portuguesa, área que tinha sido considerada prioritária.

Em vez de um projecto científico para fazer investigação, era um projecto de infraestrutura e de **serviço** à comunidade.

Após dez anos de diversas formas de financiamento e de bastante trabalho realizado, encontramo-nos numa situação de transição e de reflexão que tanto pode ser o início de uma nova fase da Linguateca como corresponder à sua conclusão.

Urge assim fazer um balanço de todo o processo e das várias fases e intenções que tivemos ao longo do tempo. Faço-o em meu nome pessoal porque fui a única que assisti e liderei este projecto desde o início, mas com o apoio de muitos e tomando em consideração todo o retorno recebido ao longo dos anos, quer dos muitos colaboradores quer da comunidade em geral, além de colher os frutos do encontro de reflexão pública em Aveiro em Setembro de 2008.

Outros textos ou apresentações sobre diferentes fases da Linguateca e sobre eventuais diferentes tónicas postas ao longo do tempo nas várias actividades podem ser consultados no catálogo de publicações da Linguateca. Saliento aqui como especialmente representativos de fases diferentes os seguintes textos (Santos, 2000; Santos, 2002b; Santos e Costa, 2005; Santos, 2007a), que serão brevemente resumidos na secção 3.2. Os vários relatórios anuais ou "finais" da Linguateca permitem dar outro tipo de visão complementar, mais concreta, cf. Santos (2003a), Santos (2005), Santos

(2006b) e Costa (2008).

2 A concepção: missão, estrutura, e ponto de partida

A Linguateca surgiu como uma forma de contrabalançar, ou resolver, muitos dos problemas ou limitações identificados durante o período da escrita do contributo para o livro branco (Santos, 1999b), há mais de dez anos, e que serão aqui repetidos esquematicamente.

Esse texto inicial, relativo à área como um todo, e de conteúdo essencialmente programático, foi uma das tarefas do projecto *Processamento Computacional do Português*[1]. Os pontos nele focados não eram para ser resolvidos na sua totalidade, ou mesmo abordados, em dez anos e por um projecto em rede. Contudo, estou convencida de que foi a nossa tentativa de não perder mais tempo e de começar logo a fazer o que era possível, ainda no âmbito do próprio projecto Processamento Computacional do Português, que levou à aprovação da Linguateca[2] nos anos que se seguiram.

É claro que os objectivos da Linguateca como projecto tiveram de ser mais concretos e realistas, embora desenhados e motivados pelos problemas que queríamos resolver e pelas metas que queríamos atingir, directa ou indirectamente. De qualquer maneira, faz todo o sentido utilizar os pontos mencionados em Santos (1999b) como uma bitola para comparar a actividade e os resultados obtidos, desde que nunca se esqueça que esse texto era dedicado à comunidade e não apenas aos membros de um projecto futuro que se viria a constituir.[3]

Vejamos então o que esse texto dizia. Antes disso, contudo, importa recordar e insistir no seguinte ponto: a área discutida e equacionada correspondia ao processamento da nossa língua e não à engenharia da linguagem em geral, veja-se Santos (1999a), o que veio a ser um dos principais cavalos de batalha da Linguateca.

Santos (1999b) mencionava as seguintes condições necessárias a um progresso significativo na área do processamento da língua portuguesa (note-se que, por conveniência da exposição, a ordem foi invertida em relação à original):

1. Transparência, participação e colaboração de todos
2. Desenvolvimento de aplicações relacionadas com o trabalho de todos os dias no sector da informação
3. Ligação da investigação fundamental com as tecnologias
4. Dinamização dos métodos empíricos
5. Serviços de desenvolvimento de recursos e ferramentas partilháveis (serviço de tradução, serviço de terminologia, rede de fala, rede de processamento da língua escrita)
6. Avaliação e controlo de qualidade em relação ao português
7. Disponibilização de recursos (nas suas múltiplas vertentes)
8. Definição do processamento do português como área prioritária

Passamos então a indagar se a Linguateca contribuiu algo para cada um destes pontos, tendo em consideração, repito, que a Linguateca foi desde o início definida como um projecto de **serviço à comunidade**, com a preocupação de não competir mas sim favorecer os actores existentes e futuros.

Mas, para o leitor incauto, convém primeiro indicar muito brevemente os pressupostos e estrutura inicial da Linguateca, ou seja, a sua espinha dorsal, antes de discutir a sua actuação e resultados.

A Linguateca, como um projecto de serviço e de apoio, foi idealizada, não através da contratação de investigadores, mas sim de "contratados" com tarefas específicas de manutenção, informação e apoio aos utilizadores, para fazer o que pomposamente se pode chamar "transferência de tecnologia" dos grupos (universitários, académicos) para o mundo exterior. Daí surgiu o conceito de **pólos** (da Linguateca), localizados em grupos ou ambientes a que faria sentido ajudar a disponibilizar o trabalho e reforçar a actividade.

Desde o início, a missão da Linguateca anunciou-se[4] como:

- facilitar o acesso aos recursos já existentes, através do desenvolvimento de serviços de acesso na rede, e mantendo um portal com informação útil,

[1] Financiado pela Agência de Inovação – organismo de financiamento português –, iniciado a 15 de Maio de 1998 no SINTEF, com a duração de dois anos.

[2] O nome *Linguateca* apenas surgiu em 2002. Do ponto de vista formal, o projecto aprovado em 2000 tinha o nome *Centro de Recursos – distribuído – para o processamento computacional da Língua Portuguesa, CRdLP*.

[3] Convém além disso esclarecer que, durante a escrita desse texto, não havia a mais remota previsão de que isso viria a acontecer, pelo menos da minha parte.

[4] De facto, esta formulação, patente na página inicial, foi pela primeira vez publicada, com algumas diferenças irrelevantes, a 9 de Agosto de 2000, como é possível verificar através do projecto Internet Archive (http://web.archive.org), ainda com o URL de www.portugues.mct.pt. A versão exacta, *ipsis verbis*, apareceu a 18 de Novembro de 2004.

- desenvolver, de forma harmoniosa, em colaboração com os interessados, os recursos considerados mais prementes,
- organizar avaliações conjuntas que envolvam a comunidade como um todo.

Assim, e ao contrário de um projecto de investigação, a nossa actividade – ou pelo menos o fundamento do nosso financiamento – repartiu-se (ou repartir-se-ia, conforme o plano) fundamentalmente entre:

- a formação de pessoal especializado em gestão, criação, disseminação e avaliação de recursos;
- o assegurar dos serviços básicos de repositório, distribuição e catálogo, de forma distribuída;
- o desenvolvimento de recursos públicos, em especial, recursos para avaliação ou calibragem;
- a manutenção do contacto e da comunicação entre os vários actores e clientes dos nossos serviços;
- a organização de avaliações conjuntas em torno de áreas chave.

Como será debatido na secção 3, de facto a Linguateca acabou por fazer muitas outras actividades não previstas inicialmente no seu desenho.

Passo então a considerar cada um dos pontos do documento original:

2.1 Transparência

A transparência foi, decididamente, uma das normas da Linguateca, embora uma questão fundamental, a da escolha dos pólos, tenha acontecido de uma forma quase aleatória, à medida que as pessoas se aproximavam de nós e se prontificavam a colaborar.

Uma das restrições (ou sugestões) que tinham sido impostas (ou recomendadas) no início era a da distribuição geográfica dos pólos, de forma a combater ou evitar a demasiada concentração de esforços num único local.

Também, do ponto de vista formal, houve ou havia restrições (inultrapassadas) no estabelecimento de pólos no estrangeiro ou em instituições privadas – o que nunca, contudo, impediu a cooperação e a formação de pólos informais, como foi o do VISL em Odense e o do COMPARA em Lisboa, ambos desde 2000.

Outra questão importante – que me parece agora explicar porque muitos grupos ou instituições não tentaram sequer obter um pólo da Linguateca – tinha a ver com a nossa filosofia de disponibilização pública dos recursos. Com efeito, fomos igualmente claros em afirmá-la, na página inicial da Linguateca, através das seguintes linhas mestras:

- Total abertura: Todas as actividades e trabalhos desenvolvidos pela Linguateca são públicos.
- Disponibilização livre: Os autores de recursos serão remunerados ou compensados de forma a não serem lesados, mas a Linguateca não se destina a desenvolver ou apoiar o desenvolvimento de recursos proprietários, mas sim a criar condições para a existência de recursos bons e gratuitos para a língua portuguesa.

Infelizmente, grande parte dos grupos na área não partilhavam ou partilham desta atitude.

Não obstante todas estas considerações, é inegável que o processo de constituição dos pólos dependeu em muitos casos da sorte, de os contactos terem sido feitos na altura certa, de as pessoas terem falado e de se terem entendido. Por isso, se a Linguateca for reaberta ou continuar, parece-nos mais correcto que todos os pólos sejam criados por concurso (aberto).

Não consideramos contudo que a primeira fase da Linguateca, por ter sido criada à medida das oportunidades que se ofereciam e dando total liberdade aos pólos – desde que com a filosofia de criarem recursos e avaliação para a comunidade – tenha sido errada ou demonstrado falta de transparência. Como é muitas vezes apontado, excesso de planeamento é geralmente sinónimo de falta de inovação (Chubin e Hackett, 1990), e ao podermos inovar, com base no material humano e tecnológico oferecido por cada pólo, fizemos muito mais do que seguir um plano rígido.

2.2 Trabalho de todos os dias

Esta é uma questão possivelmente genérica demais para ter uma concretização fácil, mas, se considerarmos que os trabalhadores nos sectores dos serviços (em que incluímos, aliás, os investigadores e desenvolvedores na nossa área) todos os dias escrevem, publicam, mandam mensagens de correio electrónico, procuram na rede e publicam na dita, além de mandarem mensagens pelo telemóvel e participarem em blogues e outras novas tecnologias, temos naturalmente de reconhecer que a actividade da Linguateca, embora com esse objectivo último, está longe de ter conseguido algum impacto, se excluirmos o círculo reduzidíssimo daqueles que pertencem ou comunicam com a Linguateca no âmbito do seu trabalho.

Assim, embora tenhamos, na medida das nossas possibilidades, apostado na promoção concreta do português através de

- sugestão de normas de redacção em português

- formas de referir publicações em língua portuguesa
- sugestões de terminologia e de desenho de sítios
- variadas intervenções em fóruns internacionais e nacionais sobre as diferenças e o respeito pela língua portuguesa
- localização e tradução para português sempre que necessário ou apropriado

não podemos considerar, de forma alguma, que esta missão – a de termos influenciado o trabalho de todos os dias das pessoas que usam o português – esteja próxima de ser cumprida.

Muito pelo contrário, cada vez mais somos instados por todos a render-nos à evidência de que o que é "internacional", isto é, escrito em inglês, é bom, e o que é nacional, isto é, escrito em português, é medíocre...

Assim, embora uma das palavras de ordem da Linguateca tenha sido a **portuguesificação**[5], demasiado ainda se encontra por fazer.

De facto, penso mesmo que estamos pior do que estávamos na altura do começo da Linguateca. Uma das convicções cada vez mais enraízadas nas camadas mais jovens – devida à forma como as agências de financiamento definem a qualidade – é que os melhores escrevem em inglês e os piores em português, o que leva naturalmente a que isso infelizmente aconteça.[6]

Alguns exemplos que demonstram claramente essa infeliz tendência são:

- o PROPOR – a conferência internacional sobre o processamento do português, com uma comissão de programa maioritariamente de lusofalantes, que desde 2003 é em inglês[7]
- a forma de avaliar os investigadores em Portugal e no Brasil: através de publicações "internacionais", mas esquecendo que o português – uma língua falada como língua materna, ou pelo menos oficial, nos cinco continentes – é uma língua internacional por excelência!
- a língua das teses e das defesas das mesmas em Portugal, que cada vez mais é o inglês em vez do português
- a língua nos sítios na rede dedicados ao processamento da língua, no Brasil e em Portugal, que cada vez mais é o inglês em detrimento do português

Veja-se, a este propósito, o valioso contributo de Gomes de Matos (1992) argumentando a favor do direito de ler e escrever na própria língua em ciência.

Por isso, parece-me evidente que a Linguateca tentou lutar contra a corrente mas que cada vez menos o português é a língua usada (ou apreciada) no local de trabalho de todos os dias.

2.3 Ligação da investigação fundamental com as tecnologias

Esta é uma atitude, mais do que uma medida: Achamos que nesta área não faz sentido uma separação, mas sim uma inter-relação entre desenvolvimento de sistemas e investigação com os mesmos.

Tentámos seguir sempre essa directiva, aliás pondo grande ênfase na questão da avaliação em tarefas práticas.

Contudo, pode ser que a linguística teórica e a informática teórica nos tenham ignorado sobranceiramente, como projecto aplicado e ateórico, e nesse aspecto a nossa intervenção tenha sido nula.

Em suma, é bastante possível que tenhamos nós mais teorizado sobre a nossa prática do que os teóricos tenham praticado graças à nossa actividade.

Não me parece, em resumo, que a Linguateca tenha de alguma forma intervindo neste aspecto, para além da sua própria prática. Que valha pelo menos o exemplo: insistimos sempre no estudo detalhado dos fenómenos da língua que poderiam estar subjacentes a um dado resultado, ou desempenho, em vez de nos ficarmos por simples medidas quantitativas deste.

2.4 Dinamização dos métodos empíricos

Neste ponto, pelo contrário, penso poder afirmar que a Linguateca contribuiu indiscutivelmente para esta dinamização, quer através da sua actividade quer através da criação de recursos que tornassem os métodos empíricos possíveis na prática.

Neste momento, na área do processamento do português, há muito mais avaliação (através de métodos empíricos) e muito maior consciência desta.

Contudo, muitas das medidas que preconizei estão longe (se calhar ainda mais longe) de serem uma realidade, senão veja-se:

> Obrigar a que todos os projectos financiados publicamente tenham uma parte de

[5] E não o aportuguesamento, ou seja, ir buscar coisas (ideias, técnicas, ferramentas) lá fora e adaptá-las ao português.

[6] Note-se que eu não estou a advogar publicação exclusiva em português, mas sim um balanço entre divulgação internacional e divulgação, didáctica e documentação na nossa língua.

[7] Na altura, a justificação avançada para esta mudança foi a de que a editora Springer concedia qualidade às publicações, e exigia o inglês como língua internacional.

avaliação (ou seja, esteja descrito na proposta como avaliar, e quando), de preferência controlável independentemente (ou seja, que a avaliação possa ser repetida por observadores externos).

Certamente que, se houve algo que não correu bem, foi a forma como o financiamento dos projectos nesta área foi atribuído em Portugal durante a existência da Linguateca – e que, acentue-se, foi sempre realizado de forma totalmente independente desta.[8]

De uma forma superficial, dir-se-ia que este foi concebido como precisamente uma compensação aos actores da área com filosofias e práticas mais distantes da Linguateca, ou seja, quanto mais "afastados" da Linguateca, mais financiamento receberiam.

Parece um critério politicamente defensável, mas os resultados práticos não o são necessariamente. Sobretudo se envolvem a repetição de esforços ou o financiamento duplo de algo já existente, como é convicção minha que aconteceu não poucas vezes.

2.5 Serviços de desenvolvimento de recursos e ferramentas partilháveis

Embora uma das áreas em que a Linguateca mais tenha investido tenha sido o desenvolvimento de serviços na rede (veja-se a secção 4.3 abaixo), tal não tomou o caminho descrito no documento preparatório. Convém talvez reflectir sobre as causas ou explicações dessa diferença aqui.

Com efeito, tínhamos preconizado a necessidade ou o interesse de desenvolver as seguintes redes de recursos:

- serviço de tradução
- serviço de terminologia
- rede de fala
- rede de processamento da língua escrita

A posteriori, parece-nos que a Linguateca se tornou **a** rede de processamento da língua escrita, e que, quanto aos outros serviços, ou foram implementados de forma completamente separada ou nunca chegaram a ser uma realidade.

Convém aqui indicar que, embora a intenção inicial da Linguateca fosse cobrir e apoiar tanto o processamento da língua escrita como da falada, tal nunca se realizou, e, após uma tentativa falhada de, logo em 2000, criar um pólo associado à fala – que nunca se materializou porque não houve candidatos a essa posição – acabámos por dirigir a nossa atenção apenas para a parte escrita.

2.5.1 Tradução automática

No início da dinamização da avaliação chegámos a criar uma lista associada à tradução automática, e vários pólos da Linguateca fizeram algum trabalho na área, mas de forma de tal maneira distinta que aparentemente não chegou nunca sequer a haver colaboração:

- O pólo do Porto dedicou-se ao estudo de ferramentas já existentes e ao trabalho necessário de pós-edição, numa perspectiva essencialmente linguística ou mesmo de estudos de tradução (Sarmento et al., 2007; Maia e Barreiro, 2007).
- O pólo de Braga dedicou-se a vários problemas tecnológicos associados ao paradigma da tradução automática por exemplos, desenvolvendo ferramentas para algumas dessas tarefas (Simões e Almeida, 2007) ou estudando a tecnologia de memórias de tradução (Almeida e Simões, 2007).
- Também se pode mencionar que implicitamente a criação do COMPARA (Frankenberg-Garcia e Santos, 2002) foi decisiva para estudos de tradução envolvendo o par de línguas português e inglês,
- assim como o pólo de Lisboa no LabEL (Barreiro e Ranchhod, 2005) produziu também algum trabalho na área.

Pese embora tanta actividade, não se chegou, pelo menos até agora, a atingir um estádio em que houvesse sistemas de tradução automática envolvendo o português desenvolvidos no âmbito da Linguateca (ou com o seu apoio) e que pudessem ser usados, embora haja algumas propostas nesse sentido, e um sistema incipiente de paráfrase (que poderá ser estendido a uma versão bilingue) foi posto ao serviço da comunidade (Barreiro, 2008).

2.5.2 Terminologia

Pior ainda, pelo menos aparentemente, foi o que aconteceu com a terminologia, visto que, embora a Linguateca tivesse desenvolvido um sistema de raiz para trabalho sério na área, o Corpógrafo (Sarmento, Maia e Santos, 2004; Maia, Sarmento e Santos, 2005; Maia, 2008b), aliás com mais de 1600 utilizadores espalhados por todo o mundo, não foi aparentemente possível congregar outras pessoas relacionadas com a área de terminologia, em Portugal ou no Brasil, de forma a trabalhar em rede.

Uma possível explicação para esse facto poderá ser a de já existirem a nível internacional várias

[8] Poderia imaginar-se que um projecto concebido para a disponibilização e avaliação de recursos poderia ser envolvido ou ser-lhe pedido um parecer quanto a novos projectos na área, com vista a garantir uma sua sustentação posterior. Cabe por isso documentar que tal nunca sucedeu.

redes de terminologia envolvendo o português[9], e como tal, em vez de criar mais uma, seria útil sim produzir sistemas que ajudassem a esse trabalho. Parece-me assim que será fundamental tentar entronizar o Corpógrafo como uma ferramenta a considerar nesses ambientes internacionais, em vez de repetir trabalho e aparecer como concorrente em vez de serviço.

Uma das questões que terá nesse caso de ser equacionada é a questão da terminologia bilingue, que, embora tenha estado na agenda do Corpógrafo desde o primeiro momento (veja-se por exemplo Maia (2003) ou Maia e Matos (2008)), ainda não tem suficiente tratamento nesse ambiente. Aliás, seria de todo o interesse aproximar (em vez de afastar) os terminólogos brasileiros, com uma longa tradição de excelência na área, note-se, e tentar na medida do possível fazer terminologia científica comum nas áreas em que isso faça sentido – a linguística e o processamento computacional da língua são, na minha opinião, uma delas.

Saliente-se, contudo, que houve algum trabalho de extracção de terminologia bilingue no âmbito da Linguateca através da tese de doutoramento de Alberto Simões (Simões, 2008).

O fosso entre abordagens linguísticas e informáticas, ao contrário do que seria a minha intenção, também ocorre(u) dentro da própria Linguateca, nunca tendo havido sinergia entre os pólos de Braga e do Porto nesse domínio.

Esse fosso, aliás já discutido por ocasião do debate em 1999[10], e que tentámos reduzir durante e através da Primeira Escola de Verão, reapareceu como não resolvido, no entender de Paulo Gomes (Gomes, 2008) ou de Belinda Maia (Maia, 2008a). Convém a esse respeito relembrar que Fernando Pereira, em 1999, tinha instado para que se criassem pessoas interdisciplinares ao contrário de equipas interdisciplinares. Ainda parece haver, no entanto, muitíssimo a fazer para que esse objectivo seja atingido.

2.6 Avaliação e controlo de qualidade em relação ao português

Em relação a este ponto, penso que a Linguateca deu um contributo decisivo, tendo-se de facto transformado no serviço preconizado em 1999:

> Seria, pois, vantajoso ter um serviço público de "portuguesificação" (por oposição a aportuguesamento) da tecnologia, incumbido de organizar as conferências de avaliação e de informar a comunidade, de garantir a distribuição dos recursos, de levar a cabo ou encomendar testes de qualidade e representar o país em órgãos internacionais

A única coisa que não aconteceu foi a "representação do país", mas dado que isso seria um trabalho sobretudo político, foi certamente preferível que esse trabalho não fosse misturado com o trabalho científico e tecnológico envolvido no resto das actividades da Linguateca, e que naturalmente nos deu muito trabalho e muito prazer.

De facto, mais do que isso: a questão "país" foi sempre substituída por "língua", tendo a Linguateca sempre defendido a língua portuguesa e não a língua dos portugueses, e tendo aliás conseguido muito boas parcerias com os investigadores brasileiros[11] exactamente por ter substituído a componente nacional por uma definida em termos da língua, que nos continua a parecer ser a única que faz sentido em termos do domínio de estudo e de prática: ou seja, no que respeita ao desenvolvimento de sistemas que lidem natural e inteligentemente com o português.

Assim, a organização de avaliações conjuntas e a sua motivação foi uma das actividades mais florescentes (e também mais absorventes) da Linguateca, como será descrito na secção 4.7.

2.7 Disponibilização de recursos (nas suas múltiplas vertentes)

Historicamente, a Linguateca foi aprovada com o nome bafiento e pouco imaginativo de *Centro de Recursos - distribuído - para a Língua Portuguesa (CRdLP)*, tendo como principal actividade a criação e distribuição de recursos.

Embora tenhamos mudado o nome e dedicado muito do nosso trabalho e empenho à avaliação, naturalmente que a criação e disponibilização de recursos – assim como a sua manutenção – foi o prato forte da actividade da Linguateca, como aliás será descrito no decurso do presente artigo.

É interessante a esse respeito ver o que foi considerado relevante em 1998 e contrastá-lo com o que temos agora (na Linguateca ou na comunidade mais vasta).

Em alguns casos, a lista referia produtos razoavelmente vagos, e noutros, demasiado específicos. Senão vejamos: Não temos provavelmente terminologias, mas temos sistemas que as permitem desenvolver; não temos dicionários com subcategorização, mas temos sistemas que permitem obtê-

[9] De facto, muito anteriores à Linguateca, como é o caso da RITERM, fundada em 1988, da TERMIP, de 1989, ou da Realiter, de 1993.

[10] cuja transcrição continua acessível do sítio da Linguateca

[11] Infelizmente, exceptuando alguns casos pontuais, a Linguateca não conseguiu (ainda) atingir ou colaborar com outros países de expressão portuguesa.

los a partir de corpos; não temos dicionários entre as variantes do português, mas temos sistemas de alinhamento que os podem eventualmente criar.

A própria terminologia também evoluiu (ou o nível de ambição): Em vez de tesauros, falamos agora de ontologias; em vez de corpos alinhados, de corpos paralelos; em vez de estudos de frequência, temos serviços que nos permitem fazê-los de forma não imaginada na altura.

Embora ainda haja certamente muitos recursos que podíamos e devíamos (como comunidade) criar, houve um claro progresso e pensamos poder afirmar que o português se encontra entre as línguas do mundo com mais recursos linguísticos públicos para o seu processamento.

Contudo, atentando nas propostas adiantadas para o conseguir, reparamos que fizemos a maior parte das coisas sozinhos, ou melhor, no âmbito da Linguateca, e não através dos meios propostos, que continuam, passados dez anos, a não passar do papel:

> a obrigatoriedade de inclusão de distribuidores e avaliadores de recursos nas próprias propostas de projectos a serem financiados, de forma a que cada centro ou grupo, além das actividades de desenvolvimento, investigação, ensino e divulgação também levasse a sério os serviços de teste, verificação e fornecimento de um serviço.

Isto continua a ser uma miragem, não há qualquer controlo de qualidade e disponibilidade dos resultados dos projectos financiados, pelo menos em Portugal.

Pelo contrário, a única coisa que se nos tornou clara em relação à disponibilização é que o nosso modelo público, **tudo grátis e sem entraves**[12], é a única maneira de chegar realmente a toda a comunidade e de evitar a mesquinhez dos tempos antigos.

Assim, como descrito na secção 4.4, comprámos o direito aos possuidores comerciais de disponibilizar recursos para todos, e isso foi um ovo de Colombo em que penso que fomos pioneiros.

Já quanto à parte da postura arquivística, também mencionada no mesmo item,

> Convém também referir que seria muito útil uma postura arquivística a respeito dos recursos, ou seja, para poder distribuir e descrever os recursos, há necessidade de criação (e de uso) de estruturas classificativas (taxonomias, tesauros classificativos); assim como se devia fomentar a codificação da informação em formatos partilháveis (tais como XML, TEI), ou pelo menos bem documentados.

temos de referir que não foi um sucesso, e isto por duas razões diferentes:

A primeira, passível de autocrítica, foi não termos tentado o suficiente. A catalogação foi sempre o parente pobre na Linguateca – ou seja, os nossos colaboradores, sem excepção, deram sempre menos prioridade a actualizar os diversos catálogos[13] do que a desenvolver sistemas ou programas ou serviços.

A segunda, no que tem a ver com a questão dos padrões, correspondeu a uma decisão pensada: considerámos sempre que o conteúdo era mais importante do que a forma, e que os padrões seriam definidos ou emergiriam do uso e não da estipulação exterior. Penso que tivemos razão, e que os padrões mencionados não são mais do que um embrulho que qualquer outro grupo pode aplicar, se precisar. Assim, os nossos padrões surgiram do trabalho que fizemos, não da adopção apriorística de regras na moda.

Em contrapartida, a documentação dos nossos produtos, serviços e recursos foi considerada de extrema importância, assim como a nossa presença na rede. Sentimos que a documentação em português era necessária quer para os falantes de português quer para a nossa identidade própria de desenvolvedores de sistemas para o processamento do português (ver secção 5.7).

2.8 Definição do processamento do português como área prioritária

Este ponto da proposta era muito vago e dirigido aos órgãos de financiamento ou organizações governativas. Até pelos percalços da actividade de governação, seria difícil de implementar ou garantir por governos sucessivos. Passe pois o conteúdo demagógico, e dediquemos apenas a atenção aos pontos concretos aventados, nomeadamente a questão da continuidade, da medida do peso da língua, a criação de um fórum, e de uma comissão internacional.

A parte ínfima que foi levada à prática foi a continuidade da própria Linguateca, no sentido em que conseguimos sobreviver dez anos e não os 2-3 anos mencionados e que continuam a constituir o prazo dos projectos de investigação.

Quanto à questão da avaliação da área, provavelmente no âmbito de um observatório estatal, nada foi para a frente que envolvesse o processamento da língua, nem mesmo a estipulação de me-

[12] No início do processo, não tínhamos esta percepção. De facto, até indico "Note-se que público não significa grátis" na respectiva secção de Santos (1999b).

[13] Como será referido em mais pormenor em 5.5.1.

didas a serem efectuadas. Contudo, existem outras instituições como a União Latina ou o Instituto Camões que poderiam tratar dessa questão. E de facto existe já há alguns anos o Observatório da Língua Portuguesa[14] que aparentemente faz alguns desses estudos.[15]

Quanto à criação de um fórum, no sentido de lista de discussão, já havia – e continua a haver – o forum-lp[16], mas que infelizmente apenas veicula anúncios (muitas vezes até em inglês!) e quase nunca discussão. Das muitas listas que a Linguateca foi criando ao longo dos anos sobre temáticas mais específicas, como avaliação conjunta, por exemplo, o mesmo resultado pode ser descrito: a comunidade portuguesa e brasileira na área do PLN não gosta nem costuma discutir questões científicas ou outras nas listas.

Se o fórum mencionado era uma conferência, temos o PROPOR, e agora no Brasil o (S)TIL e cada vez mais conferências em cada país. Mas como infelizmente o primeiro é em inglês, e o segundo não é restrito ao português, parece que ainda não existe a arena certa, ou pelo menos nenhuma especialmente dedicada e que permita a comunicação ideal dos assuntos tratados. Aparentemente, as associações de linguística de Portugal e do Brasil, APL e ABRALIN, embora ambas em países de língua portuguesa, não estabelecem fóruns comuns, e por isso também não parece possível usar nenhuma delas para dedicar ao processamento da língua portuguesa em geral, em português. Também não há (ainda?) nenhuma revista só em português sobre o seu processamento, embora a *Linguamática* seja um caso em que o mesmo é acarinhado, o que é de louvar.

Com o afã de publicação, temos de nos render à evidência: as pessoas querem publicar, não discutir nem mesmo convencer. Esse tal fórum seria ideal se fosse para as pessoas discutirem questões e da discussão sair a luz. O formato de publicação e comunicação que existe nos tempos presentes (e que não é exclusivo da nossa área ou dos nossos países) não favorece nada, contudo, esse resultado...

Finalmente, a menção de uma comissão internacional era um resquício da subserviência nacional à norma: "lá fora é melhor do que cá dentro", de que me congratulo sobremaneira não ter ido avante. No caso da língua, isso parece-me trivialmente falso. Na minha opinião, já existem demasiadas comissões internacionais de qualidade duvidosa a ameaçar a nossa soberania intelectual.

2.9 Balanço em relação ao enquadramento inicial

Santos (1999b), documento publicado na rede sem pretensões e discutido em 1999, era em muitos aspectos ingénuo e pouco fundamentado, mas apontava algumas questões concretas que era preciso atacar. Passados dez anos, é possível fazer planos muito mais concretos, e também ter muito maiores ambições quanto à área.

Agora já não falta (quase) tudo, como era o caso na altura, e a comunidade do processamento do português pode, se assim o desejar, fazer avaliação de qualidade e usar ou desenvolver recursos mais complexos. Nesse aspecto, e como aliás tentarei mostrar no resto do artigo, a actividade da Linguateca foi decisiva, embora não única.

Por outro lado, o que se passou nesta década demonstrou que, se era fácil ou possível melhorar a área no que se refere à investigação, era certamente muitíssimo mais complicado fazê-lo quanto ao impacto na sociedade em geral. Nesse ponto ainda está praticamente tudo por fazer. Voltarei a este assunto na secção 7, depois de esmiuçar as razões de satisfação – e preocupação – que o balanço da própria Linguateca me suscita.

Antes disso, porém, farei uma pequena história das várias inflexões que o projecto Linguateca sofreu, provocadas por um lado pela conjuntura político-científica distinta, e por outro por várias condicionantes pessoais da equipa da Linguateca: visto que a Linguateca são as pessoas que a compõem ou compuseram ao longo do tempo, com as suas forças e fraquezas específicas e com interesses individuais distintos.

3 A evolução

Podemos identificar alguns pontos de viragem, ou de nascimento de novas actividades, em vários momentos, não necessariamente redutíveis ao historial visível.[17]

Para referência, indica-se uma lista dos pólos[18]

[14] http://www.observatoriolp.com/

[15] O "aparentemente" deve-se ao facto de, a 30 de Março de 2009, o gráfico do "Conteúdo da Internet por línguas" se referir ao ano de 2001, e o das "Línguas da População em linha" se referir a Setembro de 2002, o que abona pouco quanto ao dinamismo e correcção de informação no dito sítio. As "Línguas de maior influência", por seu turno, referiam-se a Dezembro de 1997...

[16] Lista criada a 6 de Junho de 1997 pelo então denominado grupo "Glint - Grupo de Língua Natural DI/FCT/UNL/PT", do departamento de informática da FCT da Universidade Nova de Lisboa. Na perspectiva da Linguateca, contra a duplicação de esforços, era óbvio que devíamos apoiar e ajudar, usando, esta lista, em vez de tentar com ela competir, e temo-la usado desde sempre.

[17] No sítio da Linguateca, é possível consultar quer um historial quer uma lista de encontros organizados pela Linguateca.

[18] Conforme já indicado, muitos deles são ou foram pólos "informais" por razões administrativas. Para efeitos deste cômputo, desde que exista um doutorado associado à Linguateca, considero que um pólo existe, mesmo que a sua bolsa não seja paga pela Linguateca.

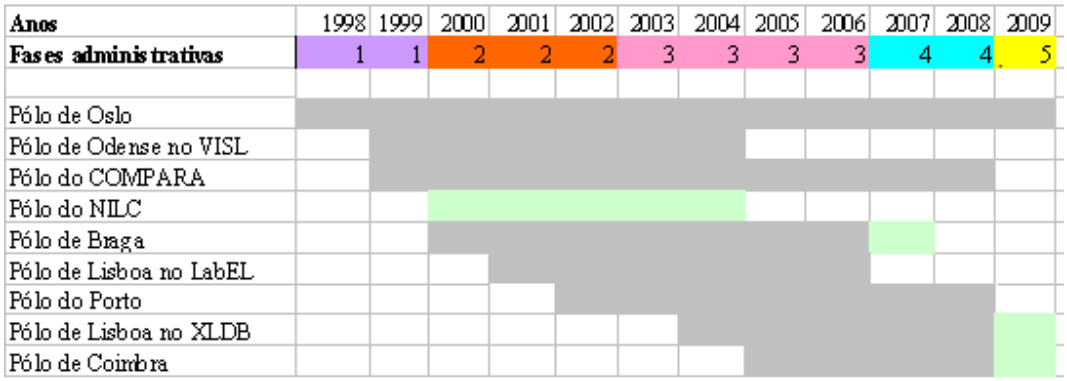

Figura 1: Actividade nos pólos, não necessariamente directamente financiada: a verde apresenta-se actividade exclusivamente no âmbito de doutoramentos

da Linguateca:

Pólo de Oslo Inicial, iniciado a 15 de Maio de 1998

Pólo do COMPARA Informalmente iniciado em 1999, formalmente transferido para a FCCN no início de 2007 e encerrado em Dezembro de 2008

Pólo de Odense Informalmente iniciado em 2000, desde 2004 apenas contando com Eckhard Bick como co-líder da Floresta

Pólo do NILC Iniciado em 2001 com o doutorado sanduíche da Rachel Aires e encerrado com a conclusão deste em 2005

Pólo de Braga Iniciado em 2000, sem pessoal afecto desde Outubro de 2007

Pólo de Lisboa no LabEL Iniciado em 2002, encerrado em Setembro de 2006

Pólo do Porto Iniciado em 2003, sem pessoal afecto desde Novembro de 2008

Pólo de Lisboa no XLDB Iniciado em Janeiro de 2004

Pólo de Coimbra Iniciado informalmente em Julho de 2005, e formalmente em Fevereiro de 2007

Além do cronograma institucional, na figura 1, e da lista dos recursos humanos com que contámos, na tabela 1, que iremos brevemente analisar na secção 3.4, podemos também mencionar actividades específicas de reunião de vários pólos num objectivo maior, e que foram fulcrais para a fertilização cruzada dos muitos ambientes distintos que compuseram a Linguateca ao longo dos tempos.

Durante os dois primeiros anos, além da preparação do documento discutido na secção 2, foram lançadas as sementes para a disponibilização dos corpos na rede (tanto o AC/DC (Santos e Bick, 2000) como o COMPARA (Frankenberg-Garcia e Santos, 2002) viram a luz do dia), e a primeira floresta para o português foi lançada, com três bolseiros em Odense (Afonso et al., 2001).

O primeiro grande acontecimento, que exigiu muito planeamento e muita discussão interna preliminar, foi o Encontro Preparatório sobre Avaliação conjunta (EPAv), com o objectivo de promover e iniciar o modelo da avaliação conjunta na comunidade do processamento computacional do português.

No ano seguinte ao EPAv, a parte de leão da actividade da Linguateca foi consagrada às Morfolimpíadas (Santos, Costa e Rocha, 2003), enquanto o pólo do Porto, o único pólo não envolvido nas ditas, dava os primeiros passos no desenvolvimento do Corpógrafo, ainda pré-baptizado "gestor de corpora" (Sarmento e Maia, 2003).

Em 2003, foi então sugerida uma expansão a nível das competências da Linguateca, que passava por ter mais formação (com a consequente atribuição de três bolsas de doutoramento), e foi integrada a área da recolha de informação, já presente desde o início do trabalho de doutoramento de Rachel Aires (Aires, 2005), através da criação de um pólo no XLDB em 2004.

Por essa altura também o CLEF (Rocha e Santos, 2007) passou a tomar um peso considerável na actividade da Linguateca, devido a estarmos nele tanto como organizadores como participantes (naturalmente, grupos ou indivíduos separados), e a sua periodicidade ser anual.

A questão das ontologias passou a ser mais uma actividade com que a Linguateca se preocupou, quer do foro geográfico quer com as ontologias lexicais criadas a partir das definições de um dicionário, o que levou à GeoNET (Chaves, Rodrigues e Silva, 2007) e ao PAPEL (Gonçalo Oliveira et al., 2008b).

A segunda actividade que congregou mais uma vez a Linguateca toda foi, contudo, o Primeiro HA-

REM, que se estendeu por quase dois anos desde o início dos preparativos até à publicação do livro a ele referente (Santos e Cardoso, 2007).

Outro acontecimento foi a (Primeira) Escola de Verão da Linguateca, que teve lugar no Porto em Junho de 2006, com todos os séniores (e alguns convidados) a disseminar o conhecimento e os recursos produzidos.[19]

Ao mesmo tempo, algumas actividades eram reduzidas ou paradas: foi o caso do serviço AnELL (Mota e Moura, 2003) no pólo do LabEL, que não chegou nunca a ter uma audiência significativa,[20] e da actividade de avaliação de tradução automática iniciada no pólo do Porto (veja-se Santos, Maia e Sarmento (2004)), que foi considerada demasiado difícil para ser continuada, com os recursos que tínhamos e as prioridades dos pólos. Também a actividade de busca inteligente, planeada como um cruzamento entre o conhecimento de terminologia e a recolha básica de informação, embora esboçada em Oliveira et al. (2005), nunca chegou a ser concretizada.

Outras ideias de projectos, ainda, não chegaram sequer a sair da fase de ideia, embora alguma publicidade lhes tivesse sido feita para obter novos colaboradores, mas em vão: um meta-dicionário (serviço na rede conjugando a consulta a muitas bases lexicais diferentes), a análise de diários às visitas ao sítio da Linguateca (e não só dos seus serviços), e interacção com fala.

Em 2006, uma nova proposta de continuação pôs a ênfase no reforço de alguns projectos com maturidade, nomeadamente o COMPARA e o HAREM (a sua segunda edição), cobrindo o resto do financiamento do programa POSC.[21]

3.1 Diferentes eixos

O modelo IRA (informação, recursos e avaliação), descrito desde o início como a trilogia fundamental da nossa actividade, foi passando a ser complementado, em novas versões da apresentação da Linguateca, com novos e variados eixos, à medida que nos compenetrávamos de tudo o que nos tínhamos comprometido a (ou tínhamos vontade de) fazer.

Senão vejamos: em Santos, Cabral e Costa (2006) ao fazer um balanço de sete anos da Linguateca, adicionámos as seguintes vertentes: manutenção de recursos, apoio, investigação (con-

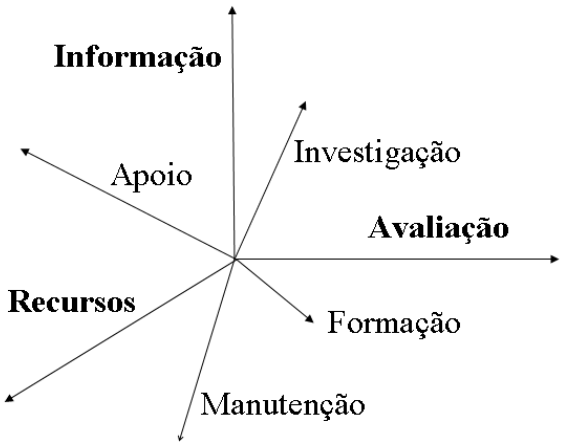

Figura 2: Eixos da actuação da Linguateca

substanciada nos doutoramentos e mestrados) e formação (relacionada com os vários simpósios doutorais e sobretudo com a (Primeira) Escola de Verão da Linguateca), veja-se a figura 2.

Ainda agora não tenho a certeza se o avançar por todos estes eixos foi uma boa ideia ou se resultou em alguma dispersão. Contudo, no âmbito da própria Linguateca, a Escola de Verão foi considerada por vários dos seus membros no encontro em Aveiro como um dos pontos altos da actividade. Possivelmente o facto de ter dado origem a – ou pelo menos influenciado positivamente – novas escolas ministradas em português: a I e II EBRalC[22], respectivamente em São Paulo e em São José do Rio Preto, e as futuras escolas que terão lugar ainda este ano de 2009, a primeira sobre "Aspectos do PLN em português", no Porto, e a III EBraLC, no Rio de Janeiro.

3.2 Formas de apresentação ao longo do tempo

Se compararmos a apresentação da Linguateca ao longo do tempo, vemos que a ênfase em catalogar e juntar os recursos acessíveis até à produção de ferramentas, sistemas ou avaliações conjuntas variou claramente.

Assim, numa leitura actual de Santos (2000), qua fazia o balanço dos dois primeiros anos de actividade, o que mais se destaca é a desproporção sobre o que, passados dez anos, fizemos em avaliação e o que pretendíamos ou imaginávamos poder fazer, em que até está mencionada a encomenda dessa actividades a actores fora da Linguateca. Assim como está bem patente a nossa esperança, depois frustrada, de incluir a fala.

Alguns pormenores interessantes mencionados, que saliento aqui, têm a ver com a preocupação de estabelecer uma metodologia (e formação) da

[19] À boa maneira da Linguateca, todo o material de ensino foi tornado público a seguir à escola, http://www.linguateca.pt/EscolaVerao2006/.

[20] Contudo, pode também interpretar-se como não ter sido totalmente implementado – de facto, outros serviços existem para o português, tais como o do VISL, http://visl.sdu.dk/, e o recente F-EXT-WS (Fernandes, Milidiu e Santos, 2009).

[21] Programa para a Sociedade do Conhecimento, activo em Portugal no período 2000-2008, http://www.posc.mctes.pt/.

[22] Escola Brasileira de Linguística Computacional

citação dos recursos criados pela Linguateca. Dada a explosão exponencial desses e doutros corpos no panorama do português, tivemos de nos render à evidência de que era quase impossível controlar ou dirigir a forma como nos citavam ou apresentavam exemplos de corpos.

Também já nessa altura pudemos apreciar que o repositório, ou seja, o serviço que iniciámos para que os investigadores que não tivessem possibilidade de o fazer tivessem uma prateleira para expor e disponibilizar os seus trabalhos na rede, não parecia muito interessante para a maioria da comunidade. Isto ainda veio a ser mais pertinente dado que a presença na rede de todas as instituições e actores passou a ser um dado adquirido, com o que aliás nos congratulamos vivamente.

Em Santos (2002b), publicado precisamente antes da escolha do nome *Linguateca*, é patente que já entrámos na espiral da avaliação conjunta, embora ainda tivéssemos a esperança de vir a ter pólos no Brasil, o que não foi nunca possível por questões políticas completamente fora do nosso alcance.

Santos e Costa (2005), por outro lado, ao apresentar a Linguateca numa revista de terminologia, põe a ênfase na publicitação dos vários recursos e projectos, constatando que, estando a infraestrutura montada, é altura de nos dedicarmos a tarefas mais complexas, de investigação aplicada. Essa previsão, e sobretudo a lista de tarefas apresentada, inspirada pelos assuntos que, na altura, se esperava que os novos doutorandos associados se dedicassem, não veio em geral a verificar-se. Mas o artigo é sintomático da fase por que passávamos (veja-se a próxima secção), que obrigava a que nos afirmássemos também como um projecto científico e não apenas de apoio e serviço à comunidade. Um foco interessante desse artigo é a descrição do levantamento feito na comunidade em 2002 sobre as áreas em que estariam interessados na avaliação, algo que foi realizado nessa altura mas nunca mais repetido ou actualizado.

Santos (2007a), por seu lado, é, até agora, o texto que melhor explica o conceito de avaliação conjunta, e a motivação para a Linguateca tomar a peito a sua divulgação e sobretudo implementação. Embora parcial porque só se refere a essa vertente, a da avaliação, foi escrito – em 2004, embora publicado em 2007 – para divulgar sem pressupor qualquer conhecimento desse paradigma de avaliação. E que muito brevemente exponho de novo aqui, para que os leitores possam compreender melhor as subsequentes referências às Morfolimpíadas, CLEF e HAREM: avaliação conjunta é a comparação do desempenho de vários sistemas com base numa tarefa comum, recursos comuns, e um aproximar de todos os interessados na área para o seu desenvolvimento e validação.

Finalmente, o presente artigo faz de novo um balanço, ao passar para uma nova fase: estou convencida de que o modelo da Linguateca tem de sofrer uma revisão substancial, e que a sua prática terá de ser mudada (ou transferida, ou encerrada) com base na reflexão que espero que este artigo possa suscitar.

3.3 Formas de apoio institucional à Linguateca (ou sua falta)

Parece-me que se deveria referir que a Linguateca não foi um projecto com um apoio estável ou com uma garantia de continuação sustentada se os seus resultados e o seu impacto fossem francamente bons – como aliás parece ser ser impossível num país da comunidade europeia ou da comunidade dos países de língua portuguesa.

Penso que, dado o financiamento e as restrições recebidas, os resultados foram bons, e a Linguateca merecia uma garantia de continuidade, mas isso não impediu a instabilidade e a total insegurança quanto à continuação do projecto em quase meia dezena de ocasiões, e aliás algumas interrupções reais de financiamento ocorridas, que não poucas vezes foram extremamente prejudiciais para os colaboradores mais jovens.

De facto, como todos os que lidaram de perto ou mesmo de longe com a nossa actividade sabem, a Linguateca materializou-se, do ponto de vista institucional, com uma sequência sempre precária e pouco reconhecida de "medidas" *in extremis* e a urgente necessidade de cumprir requisitos por vezes contraditórios de ano para ano, à medida que as fontes de financiamento foram surgindo ou mudando, assim como as regras a cumprir (de forma frequentemente inexplicável).

Se isso por um lado se deveu a diferentes governos, diferentes programas quadro e a diferentes reorganizações de tudo quanto é científico-tecnológico em Portugal e na Europa, extravasando claramente a insignificância da Linguateca e atingindo quase certamente toda a comunidade científica em todas as áreas,[23] por outro é preciso dar a ideia a quem não sabe que não fomos de forma alguma melhor tratados ou financiados do que qualquer outro projecto ou grupo em Portugal. De facto, foi elevada a percentagem de bolsas, contratos a recibos verdes, e trabalho voluntário para a Linguateca, assim como o expediente de considerar o contrato da Linguateca com o SINTEF como "investimento", de forma a garantir uma continui-

[23] Isto no que se refere ao financiamento da ciência. No que diz respeito à língua ou à cultura, ou melhor quanto à CPLP (e o seu IPLP) ou ao Instituto Camões, apesar de mais de dez anos de actividade da Linguateca, ainda não fomos reconhecidos sequer com um mero atalho nos sítios respectivos.

dade mínima (veja-se Santos (2008b) para os dados deste último).

Uma questão que foi discutida no Encontro dos 10 anos em Aveiro, mas que continua sem resolução, é exactamente que critérios de avaliação devem ser aplicados a uma iniciativa, ou organização virtual, como a Linguateca: que é ou foi concebida como um projecto de infraestrutura e não como um projecto científico.

Temos contudo e experiência negativa de em várias alturas a Linguateca ter sido avaliada (felizmente que positivamente) como se apenas de mais um projecto científico se tratasse (com critérios de número de publicações, por exemplo), o que demonstra mais uma vez um total desconhecimento ou falta de apoio dos organismos públicos que nos encomendaram a missão.

Em Costa e Cabral (2008), foram apresentados alguns indicadores sobre a Linguateca referentes a 2008, mas o estudo da verdadeira influência (ou falta dela) através de um estudo da literatura na área e áreas afins seria relevante para uma compreensão maior das consequências da nossa actividade.

3.4 O material humano associado à Linguateca

Na figura 1 apresento um quadro aproximado da ligação e trabalho efectivo dos variados membros afectos à Linguateca e pagos para tal.

Tornando a insistir na grande precariedade em que muitos elementos participaram na Linguateca, os "meses" são pois uma abstracção que se refere muitas vezes ao multiplicar e somar valores de contratos a prazo definidos à hora.

Se por um lado os mais de trinta elementos todos receberam mais ou menos formação – e pelo menos experiência – na manutenção e disponibilização de recursos e serviço continuado à comunidade, por outro as tarefas e as apetências de cada um variaram muito, conforme aliás o pólo em que estiveram envolvidas.

Se para alguns a Linguateca representou um acidente de percurso, estou convencida de que para muitos o espírito da Linguateca e o que aprenderam nela foi ou será importante para o seu futuro, e também penso que muito poucos lamentam a sua ligação.

É importante contudo salientar que escolhi fazer uma apresentação e balanço puramente pessoal – e não organizacional, como foi feito noutros casos, por exemplo em Santos et al. (2004) – e que este artigo deverá e poderá ser favoravelmente complementado pela apreciação que cada um dos séniores da Linguaetca, na sua versão pessoal, faz da sua pertença ou associação, pelo tempo que durou (no caso daqueles que já se retiraram), da vida do seu pólo e da integração ou não na Linguateca como um todo.

Diana Santos	120
Signe Oksefjell	14
Paulo Rocha	72
Tom Funcke	3
Susana Afonso	24
Miguel Oliveira	6
Rachel Marchi	18
Renato Haber	12
Alexsandro Soares	10
Rosário Silva	21
Pedro Moura	12
Anabela Barreiro	6
Luís Costa	57
Cristina Mota	22
Luís Sarmento	37
Alberto Simões	17
Luís Miguel Cabral	40
Débora Oliveira	12
Susana Inácio	50
Nuno Seco	10
Isabel Marcelino	12
Rui Vilela	26
Ana Sofia Pinto	12
Nuno Cardoso	38
António Silva	12
Ana Frankenberg Garcia	7
Sérgio Matos	12
Cláudia de Freitas	18
Hugo Oliveira	15
Pedro Martins Sousa	15
David Cruz	14
Paula Carvalho	13

Tabela 1: Colaboradores da Linguateca, por ordem de entrada (primeiro contrato), e seu contributo em meses de trabalho

Porque é preciso também relembrar que a Linguateca, mais do que a soma de todas as pessoas envolvidas, pode ser definida, estudada e explicada como a soma dos pólos, cada um deles envolvido em ambientes diferentes e com objectivos últimos diferentes.

4 Razões para satisfação e orgulho

De dez anos de trabalho em prol da comunidade, poder-se-ão naturalmente aduzir um grande número de razões para louvar e agradecer à Linguateca a sua actividade. Indico aqui as que, do meu ponto de vista, são as mais interessantes, embora não necessariamente as mais conhecidas.

Penso que em muitas destas coisas nós fomos até pioneiros a nível mundial, embora com a ressalva de que, sem a bênção da publicação interna-

cional, tal nunca será provavelmente reconhecido.

4.1 A importância da rede

Fomos dos primeiros a medir, de uma forma motivada pelo conhecimento da nossa língua, a dimensão da rede (em inglês, "Web") em português (Aires e Santos, 2002). Além disso, preocupámonos com a recolha de informação nesse contexto, em vez de usar colecções de textos jornalísticos. A primeira tese de doutoramento na Linguateca (Aires, 2005) foi pois pioneira de várias formas, e em particular pela sua intransigência determinada em recusar substitutos que não a própria rede para estudar e para desenvolver protótipos.

Também ajudámos ou incentivámos os motores de pesquisa na nossa língua e/ou cultura ao disponibilizar, e/ou ao ajudar à criação de colecções da rede disponíveis para investigação e desenvolvimento de sistemas para a língua portuguesa. A WBR-99 (Calado, 1999), a WPT-03 (Cardoso et al., 2007) e a WPT-05 são assim recursos relevantes para quem quer estudar a linguagem e a morfologia da rede em português.

Além disso temos usado cada vez mais – ao longo de uma era em que a rede cada vez mais explode em géneros e contribuições – material proveniente da vida virtual de cada um em todos os materiais de avaliação que temos tido a ocasião de criar. Assim, veja-se que, se nas Morfolimpíadas o texto da rede correspondia a menos de 10%, no Primeiro HAREM essa percentagem passou para 20% e no Segundo HAREM para 85%.[24]

Não foi também por acaso que outras teses de doutoramento se tenham concentrado em textos na rede: tanto Chaves (2008) como Cardoso (2008b), embora de forma muito diferente, lidam primordialmente com a informação geográfica na rede. Com se verá na secção seguinte, também o sistema de RAP desenvolvido na Linguateca, o Esfinge (Costa, 2005), usa a redundância da rede como um elemento principal.

Finalmente, o próprio uso da rede como recurso para outro tipo de dados, por exemplo para a compilação de corpos paralelos, também foi investigado pelo pólo de Braga desde muito cedo, como se pode apreciar em Almeida, Simões e Castro (2002).

4.2 Novos modelos de resposta automática a perguntas

Estou também convencida de que a Linguateca deu uma contribuição importante à área da resposta automática a perguntas, RAP – e não só à existência de vários sistemas e grupos interessados nessa aplicação para o português.

Com efeito, desde 2004 que somos co-responsáveis pela organização da pista de RAP do CLEF, QA@CLEF, incluindo o português, veja-se por exemplo Vallin et al. (2005) e Forner et al. (2009), e o que é um resultado indiscutível do CLEF é que já em 2007 o português foi a língua com mais sistemas participantes de RAP.

Contudo, a Linguateca também foi autora de uma proposta inovadorada de RAP colaborativa (Santos e Costa, 2007); da disponibilização de colecções sintacticamente anotadas para teste e treino de sistemas de RAP (Santos e Rocha, 2005); de um sistema desenvolvido de raiz para o português em código aberto, o Esfinge (Costa, 2005; Costa, 2006); e duma avaliação conjunta pioneira, o GikiP (Santos et al., 2009), seguido pelo GikiCLEF, em progresso neste momento.[25]

Além disso, embora indirectamente, esperamos contribuir para a existência de mais trabalhos de investigação na área ao incluirmos perguntas na colecção do Segundo HAREM, conforme explicado em Carvalho et al. (2008).

Ao contrário de muito do trabalho corrente em RAP, cuja preocupação é melhorar alguns pontos percentuais no desempenho de sistemas, sem entrar em conta com a realidade e/ou pertinência da tarefa ou com a validade linguística dos modelos empregues (veja-se por exemplo a tarefa de detecção do tipo de resposta descrita em Roberts e Hickl (2008)), a nossa actuação tentou sempre pautar-se por trazer a RAP para a realidade das necessidades do utilizador e não de uma comunidade científica específica.

4.3 Recursos realmente acessíveis

O que fizemos com o projecto AC/DC foi de facto pioneiro – colocar todos os corpos que pudemos disponibilizar acessíveis de uma maneira idêntica, para facilitar o seu uso e manipulação com um mínimo (ou nenhum) conhecimento informático (Santos e Bick, 2000; Santos e Sarmento, 2003).

Convém relembrar que na altura não havia nenhum sistema de procura ou acesso a corpos em português, e os poucos corpos existentes eram levantados em conjunto (ou seja, por "download").

Depois disso, muitas outras instituições – algumas sem sequer nos mencionar ou citar (Bacelar do Nascimento, Mendes e Pereira, 2004; Aluisio et al., 2004), outras explicitamente explicando que o nosso modelo não lhes convinha (Aluísio, Oliveira e Pinheiro, 2004) – puseram os seus corpos também acessíveis na rede.

Outros ainda criaram novos corpos e novas interfaces, o Corpus Informatizado do Português Me-

[24] No caso do Segundo HAREM, estou a contar apenas a colecção dourada, visto que a colecção do Segundo HAREM foi obtida a partir dessa e da colecção CHAVE. Para mais pormenores, ver Santos et al. (2008).

[25] Veja-se http://www.linguateca.pt/GikiCLEF/.

dieval (Xavier et al., 1998), o Corpus do Português (Davies e Preto-Bay, 2008), o Corpus Brasileiro (Berber Sardinha, Moreira Filho e Alambert, 2008). De facto, podemos agora afirmar que não existe efectivamente falta de material anotado sobre o português, embora eu ache que do ponto de vista da documentação, o material da Linguateca é ainda incomparavelmente superior – o que não significa que não possa ser melhorada.[26] Por outro lado, no que respeita à usabilidade e à experiência de interacção proporcionada ao utilizador, estamos decididamente bem atrás destes três projectos.

Não é possível, naturalmente, pronunciar-me sobre se todas estas iniciativas teriam existido na mesma sem a Linguateca, ou se, pelo contrário, apareceram como uma resposta, positiva ou negativa, à nossa actividade.

4.4 Modelos económicos

Uma questão em que a Linguateca sempre insistiu foi a de não dever haver diferença entre usos comerciais e usos académicos. Tal distinção foi, aliás, considerada um dos principais entraves à fertilização cruzada entre investigação e produtos com impacto no dia a dia.

Assim, o CETEMPúblico (Rocha e Santos, 2000) foi negociado com o jornal PÚBLICO exactamente nessa base, assim como o PAPEL (Gonçalo Oliveira et al., 2008b) e o CLASSLPPE, com a Porto Editora, o foram também. Estes casos são aliás a prova cabal de que não há uma distinção de mentalidades entre empresas e universidades. De facto, e ao contrário da tese "as companhias privadas só querem o proveito próprio, enquanto os universitários estão conscientes do seu papel social", as empresas foram em geral mais receptivas a disponibilizar do que muitos grupos ou investigadores individuais.

Talvez também seja de realçar que, mais uma vez ao contrário do que poderia ser esperado, foram sempre sistemas comerciais ou semicomerciais que venceram as avaliações conjuntas que organizámos: nomeadamente o PALAVRAS (Bick, 2000), o CorTex (Aranha, 2007) e o sistema da Priberam (Amaral et al., 2008). Não se pode, pois, partir de uma hipótese definitivamente não corroborada para continuar a defender a excelência académica por oposição à cegueira empresarial: no contexto da língua portuguesa, isto simplesmente não é verdade.

[26] Veja-se por exemplo a documentação sobre a revisão da anotação morfossintáctica da parte portuguesa do COMPARA (Inácio e Santos, 2008), que pretende indicar todas as opções tomadas em algo que é obviamente não trivial.

Tipo de texto	Abs.	Tam.	Rel.
Texto traduzido	444	723807	61,34
Texto original	258	818553	31,52

Tabela 2: Diferença entre texto original e traduzido no que se refere a *already* no COMPARA 13.1.4.

Expressão	Freq. absoluta	Freq. relativa
já	3121	2,17
já - already	811	0,56
already	916	0,59

Tabela 3: Ocorrências de *já* e de *already* no COMPARA, versão 13.1.4.: a frequência relativa é por mil palavras da língua respectiva

4.5 Corpos paralelos

Outra área em que a Linguateca muito fez foi na disponibilização e divulgação de corpos paralelos através do COMPARA (Frankenberg-Garcia e Santos, 2002) e, mais tarde, do CorTrad[27]. Que eu saiba, o COMPARA é o maior corpo paralelo revisto morfossintacticamente no mundo inteiro, e tem algumas funcionalidades únicas, tal como a procura por notas de tradução e a distribuição cruzada (Santos, 2002a). Além disso tem anotação semântica revista (Santos, Silva e Inácio, 2008), algo que também é raro, senão único, em corpos paralelos.

Ainda podemos salientar o facto de uma das primeiras análises quantitativas da interacção dos utilizadores com um corpo paralelo ter sido feita no COMPARA (Santos e Frankenberg-Garcia, 2007).

Contudo, um erro cometido no âmbito do COMPARA foi a dependência demasiada em relação a autorizações demasiado específicas de algumas editoras, o que implica (ou implicará, num futuro próximo, dependente de cada autorização) o retirar dos pares de textos respectivos do acesso público. É minha convicção agora que não deveríamos ter investido tanto trabalho (de revisão e anotação) em textos que teriam uma vida pública breve.

De qualquer maneira, noto que o DISPARA facilitou enormemente a obtenção de dados e de pesquisas num corpo paralelo: por exemplo, para obter a informação de que *already* é mais frequente em texto traduzido do que em texto original (ver tabela 2), ou de que *já* corresponde mais a *already* do que *already* a *já* (ver tabela 3), tabelas laboriosamente obtidas durante o meu doutoramento, e referidas entre outros em Santos (1995) ou Santos (2008c), basta um simples comando no DISPARA.

[27] O CorTrad é um subprojecto do projeto COMET - Corpus Multilíngüe para Ensino e Tradução, da Universidade de São Paulo, cuja disponibilização é feita através do sistema DISPARA, em parceria com a Linguateca e o NILC.

4.6 Análise gramatical

Outro dos pressupostos científicos da Linguateca, que pensamos ter sido completamente demonstrado, foi a inutilidade, e mesmo prejuízo, de focar em "POS tagging" (anotação da categoria gramatical em contexto) em vez de tentar uma análise sintáctica mais complexa. Como defendido em Santos (1999c), essa aplicação é boa para o inglês, mas pouco apropriada para línguas que, como o português, têm mais de setenta formas verbais diferentes, além de um sistema complexo de enclíticos e mesoclíticos. Claramente a ênfase no que é problemático (e fácil) na nossa língua é mais útil do que a importação acrítica de modelos criados para línguas diferentes.

É certo que o facto de termos um pólo em Odense levou a que a Linguateca favorecesse, no sentido de publicitasse, o PALAVRAS (Bick, 2000), mas não só é preciso indicar que isso se deveu ao desejo de Eckhard Bick colaborar com a Linguateca (uma colaboração que se afigurou vantajosa para ambas as partes), como não houve nem há nenhum outro sistema de análise gramatical comparável para o português, pelo menos de que eu tenha conhecimento. Por essa razão, existe de certa forma um monopólio do PALAVRAS para o processamento da língua portuguesa.[28]

Contudo, penso dever salientar que a Linguateca contribuiu para melhorar o PALAVRAS de várias formas distintas e não insignificantes: Por um lado, ao ter entrado em vários projectos conjuntos que incluíam o VISL, em particular a Floresta Sintá(c)tica (Afonso et al., 2001; Bick et al., 2007; Freitas, Rocha e Bick, 2008a), em que um dos objectivos principais era mesmo a melhoria do analisador sintáctico e das suas bases teóricas para a descrição do português real (ao congregar uma equipa de linguistas debruçada sobre os mais ínfimos pormenores), veja-se a secção 4.8. Por outro lado, a colaboração e uso do PALAVRAS em outros projectos, nomeadamente o AC/DC, o COMPARA, o Esfinge[29] e o CorTrad, levou a que fossem sendo enviados ao longo do tempo extensos relatórios de problemas ou de sugestões relativas à análise sintáctica computacional em português.

Saliente-se também que os corpos anotados no âmbito da Floresta e do AC/DC estão acessíveis publicamente (nos casos em que os detentores do material no-lo permitiram), assim como o serviço SketchEngine[30] (Kilgarriff et al., 2005), que produz uma descrição automática das propriedades gramaticais e contextuais das palavras para efeitos lexicográficos, é grátis para o português – e só para o português – porque baseado nos corpos anotados da Linguateca.[31]

Esses corpos anotados deram aliás origem pelo menos a um analisador estatístico público para o português (Wing e Baldrige, 2006).

Outro lado da nossa aposta na anotação gramatical foram as várias tentativas de discutir e/ou de centrar a atenção em muitos aspectos da análise da língua portuguesa ainda pouco explorados, ilustrados por Santos e Gasperin (2002), Afonso (2003), Santos (2004), Afonso (2004) ou Inácio, Santos e Silva (2008).

Refiram-se também as várias acções pedagógicas e de explicação dos vários conceitos envolvidos, que foram realizadas em várias ocasiões (Santos, 2006a; Santos, 2008a) além da constante ajuda aos utilizadores dos vários projectos envolvendo anotação gramatical.[32]

Finalmente, a nossa "Bíblia florestal" (Freitas e Afonso, 2008) não pode deixar de ser referida como um dos trabalhos mais extensos e completos, baseados em texto, criados nos últimos tempos sobre a análise sintáctica do português, e cobrindo, além disso, as duas variantes da língua.

4.7 Avaliação conjunta

Quanto à avaliação conjunta, foi a área em que decididamente houve mais progresso no processamento computacional da língua portuguesa nestes dez anos:

Passámos de uma total ausência e desconhecimento desse paradigma até à implantação forte do modelo em (quase) toda a comunidade, e com o consequente reconhecimento da necessidade e utilidade de novas iniciativas.

Para isso a Linguateca foi absolutamente fundamental, desde a formação e divulgação até à concepção de iniciativas de reconhecido valor internacional e com pressupostos originais e únicos.

Visto que temos um livro expressamente dedi-

[28] Esse "monopólio" não é, contudo, obra da Linguateca: o PALAVRAS tem sido empregue por quase todos os grupos de PLN no Brasil ou Portugal, sem qualquer relação com a nossa actividade.

[29] Neste último caso, o PALAVRAS é usado apenas para a parte da referência anafórica, ver Cabral, Costa e Santos (2007).

[30] http://www.sketchengine.co.uk/

[31] Pelo menos foi essa a combinação feita com Adam Kilgarriff e Eckhard Bick quando nos foi pedida autorização para usar o CETEMPúblico e o CETENFolha. Não me pronuncio aqui sobre novas licenças e/ou formas de aceder a esse serviço que não incluam nem sejam baseadas em material da Linguateca, mas insisto em que a Linguateca não tem quaisquer objecções a que o material por nós criado seja usado por empresas ou para fins comerciais.

[32] Esta é uma actividade que é de certa forma invisível, a não ser para aqueles que a recebem directamente, mas que pode corresponder a uma diferença significativa em termos da utilidade para o exterior dos corpos e recursos disponibilizados. Pensamos que esta característica é especial da Linguateca, e que tal não acontece com a maior parte dos outros recursos ou serviços na rede, embora não tenhamos, naturalmente, dados objectivos para o afirmar.

cado a esse paradigma (e incluindo os participantes nas Morfolimpíadas) (Santos, 2007b), assim como dois outros livros referentes às duas edições do HAREM, Santos e Cardoso (2007) e Mota e Santos (2008), não me nos vou alongar aqui.

Gostava contudo de salientar três traços importantes desta actividade que nem sempre são óbvios para quem está de fora:

- a criação e disponibilização pública de ferramentas e serviços de avaliação (Seco et al., 2006; Gonçalo Oliveira et al., 2008a; Cardoso, 2008a);
- a documentação e reflexão sobre os recursos, também públicos, de avaliação (Santos e Barreiro, 2004; Barreiro e Afonso, 2007; Cardoso e Santos, 2007);
- a congregação de comunidades até aí inexistentes mas que se dedicam a uma mesma tarefa (Santos, 2007a).

Além disso, convém também apontar que o ReRelEM (Freitas et al., 2008; Freitas et al., 2009), a tarefa de detecção de relações entre entidades mencionadas proposta no Segundo HAREM, ao conseguir um cruzamento entre a detecção automática de referência anafórica, tal como por exemplo analisada pelo MUC (Chinchor e Robinson, 1998) ou pelo ARE (Orăsan et al., 2008) e a detecção de relações em texto típica da extracção de informação constitui um desafio original, embora com parecenças com o ACE (NIST e ACE, 2007), que coloca o português entre as línguas que desbravam o processamento da linguagem natural.

4.8 A floresta mais complexa do mundo?

Embora a Floresta Sintá(c)tica não tenha tido o sucesso ou impacto – em termos de utilizadores – que esperaria, penso que foi um projecto inovador e de grande qualidade que possivelmente criou uma das primeiras florestas com informação sintáctica complexa para qualquer língua.

Porque este me parece um caso paradigmático de falta de impacto na comunidade apesar de um esforço considerável para o contrário, refiro que a equipa tentou "tudo" para congregar o máximo de actores à volta dela, senão vejamos: i) apelámos ruidosamente no início do desenvolvimento da Floresta para que fosse um projecto de colaboração entre toda a comunidade, a quem pedíamos para sugerir e prover novos textos e novos analisadores automáticos; ii) temos feito ao longo dos tempos sempre muita divulgação em departamentos de linguística e de computação no Brasil e em Portugal; iii) temos insistido em que se pode obter dados mais simples (tal como sintagmas nominais não complexos) para (avaliar) tarefas que apenas precisem de análise superficial; iv) a Floresta existe numa quase dezena de formatos diferentes "ao gosto do freguês" (Vilela et al., 2005), e com variada informação, semântica, anafórica, de discurso, etc. (criada pelo VISL), (v) finalmente, está integrada em diversos ambientes de processamento internacionais, tal como o NLP toolkit[33], assim como foi usada em avaliações conjuntas internacionais, como o CoNLL.

Muitas das opções tomadas e das ferramentas desenvolvidas no âmbito da Floresta também me parece terem sido originais: Por exemplo, o Picapau (Haber, 2001) está bem à frente dos sistemas desenvolvidos para lidar com florestas, como aliás se vê pela resenha e descrição feita em Lai e Bird (2004), que infelizmente também não menciona o Águia (Santos, 2003b).[34]

Convém reflectir sobre a Floresta Sintá(c)tica e sobre a pertinência da sua criação: O que é certo é que existe um recurso, por enquanto muito pouco explorado, mas que permite uma enorme riqueza de estudos e pesquisas ainda por estabelecer. A que ponto é que tal riqueza seria necessária em 2000 (ou agora)? Deveríamos antes ter começado pelas coisas mais simples? Isto é algo que tem sido bastante discutido pela comunidade que nos cerca.

A minha opinião é que teria sido redutor não tentar ambos os caminhos, apostando assim no servir o máximo de público e de colaboradores interessados, embora não desprezando outras formas de produzir recursos menos ambiciosos. Veja-se uma discussão inicial sobre o assunto em Inácio e Santos (2006), contrastando a revisão do COMPARA com a criação da Floresta. Para outras achegas para o debate em torno da Floresta consulte-se as apresentações de balanço no Encontro "Um Passeio na Floresta Sintáctica", e os novos rumos e interfaces do projecto (Freitas, 2008; Freitas, Rocha e Bick, 2008b).

4.9 Publicar e catalogar em português

Uma das questões mais óbvias que se nos deparou no nosso trabalho interno de todos os dias foi a falta de qualidade dos sistemas de gestão de referências "internacionais" para lidarem com os falantes, e autores, de língua portuguesa, o que levou a que acabássemos por ter de gizar de raiz um sistema para garantir esse (algum) controlo de qualidade, o SUPeRB (Cabral, 2007; Cabral, Santos e Costa, 2008).

Em paralelo, a nossa experiência convenceu-nos

[33] http://www.nltk.org/

[34] É que, como aliás voltarei ao assunto mais à frente, na minha opinião também existe, na comunidade de língua inglesa, o preconceito de que "o que não está ainda feito para o inglês, não existe", mesmo que publicado em inglês.

também de que a actualização manual de um sítio, sem ajuda automática, é muito pouco eficiente e possivelmente condenada ao insucesso (veja-se, por exemplo, a discussão em Pekar e Evans (2007) sobre os catálogos na rede), e que o ideal são sistemas supervisionados em que o processamento automático é depois validado por especialistas: aliás uma opção que nos parece fazer sentido em quase todas as áreas de PLN.

Assim, ao mesmo tempo que tentávamos aplicar a tecnologia e o conhecimento do processamento da nossa língua na nossa actividade quotidiana, nomeadamente na catalogação (das publicações) da área, desenvolvemos um serviço e um sistema que poderia extravasar claramente a área da engenharia da linguagem e ser utilizado por todos os membros da comunidade científica lusofalante, ou seja, um SUPeRBibliotecário desenvolvido de raiz para o português mas com consciência e conhecimento do mundo da publicação em inglês e noutras línguas (por agora, apenas europeias).

Este sistema, além de ser subjacente ao catálogo de publicações da Linguateca (na área), e às variadas páginas de publicações de cada subprojecto (criadas automaticamente), foi usado no desenvolvimento e preparação dos vários livros e artigos desenvolvidos na Linguateca, e encontra-se, quer como serviço, quer como programa em código aberto, acessível publicamente.

4.10 A contribuição das Morfolimpíadas

Parece-me importante retirar do esquecimento as Primeiras Morfolimpíadas para o português, porque, embora não tenha havido seguimento nem aparentemente resultados baseados em estudos sobre os recursos tornados acessíveis, várias coisas ficaram claras:

Por um lado, a existência de fortes divergências teóricas e de diferente importância dada a diferentes fenómenos entre grupos que desenvolveram ou desenvolviam sistemas de análise morfológica.

Por outro lado, uma medição concreta – e extremamente significativa – das diferenças em relação à atomização praticada por cada grupo (Santos, Costa e Rocha, 2003).

Mais uma vez penso que estas medidas foram as primeiras para qualquer língua, embora naturalmente outras medidas e outros problemas tivessem sido privilegiados para o alemão (Hausser, 1996), a língua em que a primeira avaliação conjunta relacionada com morfologia computacional foi levada a cabo. Basta, contudo, reconhecer que esta última língua tem o problema dos compostos para se compreender que outras questões e outras medidas fazem sentido nas duas línguas.

Finalmente, parece-me que também ficou claro que, por ser uma tarefa demasiado teórica, ou seja, dependente de uma separação arbitrária entre níveis ou estratos de língua, muitas das opções ficaram por avaliar, visto que não se encontravam inseridas numa tarefa concreta com resultados consensuais, independentes do modelo teórico.

5 Razões para preocupação

Não gostava contudo de terminar este balanço sem indicar que também houve muita coisa que correu mal, ou que poderia ter corrido melhor. Apresento aqui estes variados pontos para ajudar a fazer não só uma apreciação justa da nossa actividade, como para permitir a outros ou a nós, a começar de novo, não cometer os mesmos erros ou pelo menos ter logo em conta os riscos apontados.

Os quatro primeiros itens têm a ver com a aceitação ou relação da Linguateca com o seu contexto, e podem pois considerar-se do foro sociológico. O quinto ponto refere críticas que nos foram feitas e com que concordo total ou parcialmente, ou que pelo menos considero importante reconhecer a sua existência. Os últimos pontos discutem questões reconhecidamente difíceis mas com cujo tratamento não me considero, de qualquer maneira, totalmente satisfeita.

5.1 Pouco impacto

Atingimos muito poucas pessoas das que poderíamos ter atingido. A grande maioria das pessoas relacionadas com a língua portuguesa ou com a cultura portuguesa nunca ouviu falar da Linguateca. Isso reflecte-se tanto em alunos de doutoramento em Portugal e Brasil como em pesquisadores brasileiros ou portugueses em áreas centrais ou próximas. Ainda agora nos aparecem pessoas que "encontraram o nosso sítio por acaso".

Se isso de certa forma constituiu uma escolha nossa, por termos definido como base de utilizadores (e beneficiários) as pessoas que trabalhavam em ou com o processamento do português (ou seja, a área do PLN, da engenharia da linguagem ou da linguística computacional), e não com a área da língua portuguesa em geral, parece-nos de qualquer maneira que o nosso impacto (e consequente utilidade) deveria ter sido maior.

Da mesma forma, em áreas em que a nossa actividade poderia ter abrangido muito mais gente, como é o caso da publicação científica em geral, e em particular a criação de listas bibliográficas em português ou incluindo correctamente autores de língua materna portuguesa, aparentemente ninguém sabe que fizemos algo que lhes pode ser útil, e que está público. Daí existirem muitos e variados projectos e iniciativas, até de criar bibliografias relacionadas com a área (por exemplo de linguística), que poderiam beneficiar de interacção,

colaboração e troca de dados e das próprias ferramentas desenvolvidas, mas que não utilizam aquilo que oferecemos ou poderíamos oferecer.[35]

Isto demonstra que a colaboração com outras instituições e o reuso de materiais ou trabalho feito por um dado projecto é algo muito mais complexo e exige muito mais atenção do que ingenuamente supusemos.

5.2 Pouco reconhecimento

Uma questão que está relacionada com o pouco impacto e que talvez contribua para ele mesmo é a falta de reconhecimento público aos serviços ou recursos desenvolvidos ou providenciados pela Linguateca.

Penso que não é exagero dizer que mesmo as pessoas que têm bom conhecimento da Linguateca não fazem em geral qualquer esforço para a citar como deve ser, pese embora a nossa continuada insistência em providenciar modelos e até explicitamente indicar como os recursos ou o nosso trabalho devem ser citados. De facto, temos na lista de perguntas já respondidas a informação de como citar cada recurso, assim como muitas vezes na própria página do dito recurso. No entanto, a maior parte das pessoas, se citam, dizem simplesmente "o corpus do Público" (ou "da Folha") ou até os "corpos da Linguateca".

Mesmo as pessoas dentro da Linguateca demonstram o espírito "fora é melhor", porque dá publicação internacional, como se pode ver pela apresentação do Mário J. Silva no encontro que fez um balanço da Linguateca passados dez anos (Silva, 2008b). Segundo ele, o trabalho feito pela Linguateca no CLEF foi muito mais útil e importante que o por exemplo do HAREM, mesmo que a participação de grupos de processamento da língua portuguesa tenha sido mais reduzida[36] e a influência e qualidade do trabalho feito em relação ao português seja incomparavelmente menor[37], dado que a exposição internacional é muito superior no primeiro.

Mas, se esse espírito continua na comunidade do processamento do português, por definição impede que o português atinja a maioridade científica, o que era exactamente uma das intenções da Linguateca: demonstrar que, para o processamento da língua portuguesa, os próprios membros da comunidade que conheciam a língua como sua língua materna eram naturalmente os melhores para essa tarefa.

De facto, a questão do português na comunidade internacional é de alguma forma interessante problematizar: não só considero (Santos, 2007c) bastante pernicioso para o próprio PLN em geral, como disciplina que não haja investigação feita de novo para outras línguas – em particular a nossa – como é muito mais fácil publicar dados empíricos errados ou mal interpretados quando a comissão de programa não percebe a língua. Além disso, convém não esquecer que a maioria dos nossos colegas anglofalantes têm arreigada uma concepção completamente errada, na minha opinião, da área, e que se traduz no seguinte: "todas as inovações começam no inglês", donde a história da área faz-se com base sempre, ou quase sempre, na história da cultura anglo-americana.

No entanto, se os portugueses e brasileiros continuarem sem citar nem mencionar os seus pares na comunidade do processamento do português, e se projectos como a Linguateca não receberem a menção que deveriam ao ter contribuído para o trabalho descrito, está-se a perpetuar essa percepção na comunidade internacional, e na da língua portuguesa.

5.3 Falta de confiança?

Embora a Linguateca tenha dito desde o primeiro dia que queria servir a comunidade, a nossa oferta de disponibilizar os corpos de outras instituições foi recebida com desconfiança (quase) total, e essas instituições foram desenvolver e criar as suas próprias soluções (com o seu próprio financiamento ou com financiamento público), o que teria sido muito mais bem empregue em parceria connosco em vez de contra nós.

Com efeito, nós oferecemo-nos para disponibilizar todos os corpos de português existentes (através do projecto AC/DC). Contudo, muitos projectos para fazer exactamente isso foram iniciados e levados a cabo depois. Dado que nós oferecíamos a tecnologia e o nosso saber-fazer, e muitas dessas instituições até eram académicas e não especialmente interessadas em tecnologia ou disponibilização, é difícil compreender a rejeição, ou ignorância voluntária, dessa oferta.

Outra dessas manifestações é a procura de uma dada ferramenta e/ou serviço, que depois, ao descobrirem que não existe para a língua portuguesa, ou pelo menos não na Linguateca, acaba numa proposta de projecto que, regra geral, não inclui como colaboração ou parceria, ou sequer consultoria, a Linguateca.

Não seria melhor para todos se também se acon-

[35] Veja-se a título de exemplo a Bibliografia Corrente de Linguística do Português, http://dupond.ci.uc.pt/celga/, com apenas dezassete entradas de linguística computacional em Abril de 2009.

[36] Na pista geral do CLEF e no GeoCLEF, em cinco anos e portanto cinco edições participaram apenas quatro grupos diferentes, brasileiros ou portugueses, entre os mais de quarenta. No HAREM participaram vinte em duas edições.

[37] Como pode ser facilmente apreciado, sendo preciso discutir e chegar a consenso com uma miríade de co-organizadores encarregados das outras línguas.

selhassem, ou perguntassem a nossa opinião sobre uma possível colaboração ou participação no desenho dos requisitos, em vez de apenas nos utilizarem como bibliotecários especializados? Mais uma vez, penso que essa forma de proceder não é a melhor para a comunidade como um todo, porque dá prioridade aos interesses específicos de um dado grupo.

Outra possibilidade aventada para explicar este comportamento é a questão do protagonismo. É melhor fazer as coisas sozinho, para receber todos os louros, e o reconhecimento de ser primeiro ou original, do que em colaboração com outros, aliás porque o financiamento é por competição.

De facto, uma das coisas que se tornou mais clara para mim é que muitas pessoas preferem independência a colaboração, e que não são movidas por um desejo de avançar a área como um todo, mas sim de se tornarem os líderes incontestados num determinado nicho ou sub-área.

Será preciso reflectir se esta atitude é saudável ou se é preciso reforçar a interdependência ou, pelo contrário, proceder a uma distribuição de feudos por diferentes actores para estimular o progresso.

De qualquer forma, a única afirmação que é indiscutível é que, mesmo sempre nos apresentando como um serviço, muitos houve que não quiseram partilhar a fama ou os trabalhos connosco.

Outra questão que é preciso mencionar e que é de grande importância tem a ver com o facto de a Linguateca ter sido um projecto iniciado por Portugal e de nunca se ter conseguido (ainda?) pôr de pé os mecanismos formais para criar pólos no Brasil, assim como uma estrutura paralela ou geminada. Isto faz ou fez com que de facto seja muito mais difícil estabelecer projectos comuns com grupos brasileiros e/ou sobretudo obter financiamento para tal.

Ora exactamente para aproveitar o facto de que em português nos entendemos seria essencial promover um apoio, por exemplo, à participação em avaliações conjuntas especialmente promovidas para estimular o progresso do processamento do português, assim como à realização e promoção de fóruns, conferências, encontros, escolas, em português para discutir a língua e o seu processamento.

5.4 Livros difíceis de obter?

Um dos resultados mais fácil de medir objectivamente é a actividade de organização de livros no âmbito da Linguateca: quatro livros distintos sobre a actividade da Linguateca vieram à luz (Santos, 2007b; Santos e Cardoso, 2007; Costa, Santos e Cardoso, 2008; Mota e Santos, 2008).[38]

Mas, além de tal actividade se ter demonstrado muito complexa, tenho fortes dúvidas de que os resultados sejam positivos no cômputo geral: Com efeito, o objectivo de organizarmos nós próprios os livros é podermos ter o controlo total da qualidade, e aliás dos assuntos tratados. No entanto, se esses livros não receberem um canal de publicação apropriado e não forem portanto passíveis da divulgação por nós desejada, não cumprirão o seu objectivo.

Em relação ao primeiro livro, não só se revelou um processo complicadíssimo obter uma saída editorial (atrasando mais de três anos a distribuição do seu conteúdo), como a opção por uma editora comercial impediu a fácil divulgação dos textos. No segundo e terceiro casos, a opção de publicar directamente na rede, embora resultando numa divulgação muito mais rápida, diminuiu claramente o valor científico-comercial do produto, e possivelmente mesmo a sua longevidade.

Neste momento, dado que nenhuma alternativa parece ser realmente satisfatória, ainda nos encontramos num processo de reflexão no que se refere à publicação da quarta obra.

5.5 Críticas variadas

Não posso naturalmente deixar de reconhecer que muitas das críticas que nos foram feitas, aliás por ocasião do balanço dos dez anos, são justas e merecem que as reconheçamos como pontos em que falhámos.

5.5.1 Egocentrismo institucional

Uma das missões da Linguateca era a de catalogar a área, construindo um portal de entrada para tudo o que existisse na rede e pudesse ser útil ao processamento computacional do português.

Contudo, é fácil de ver que o nosso sítio (do qual se apresenta um ecrã na figura 3) está muito mais centrado na nossa actividade do que na da catalogação (Nunes, 2008). Com efeito, ao lado dos catálogos de recursos, ferramentas, actores e publicações, que reflectem ou deviam reflectir a área como um todo, temos muitíssimas outras opções para seduzir o visitante incauto ou interessado, que não vá já com um objectivo determinado.

Em primeiro lugar, damos "Acesso a recursos" da Linguateca primeiro que ao catálogo em geral, "Catálogo de recursos", e iniciamos a lista de opções no menu da esquerda pela pouca modesta apresentação (da Linguateca); depois juntamos, além dos catálogos e de informação interessante, a rubrica "Avaliação conjunta" em que também tivemos um papel fundamental.

Em segundo lugar, os itens "sistemas de procura" e "perguntas já respondidas", que são uti-

[38] Outros livros também organizados parcialmente no âmbito da Linguateca foram Almeida (2003) e Peters et al. (2008).

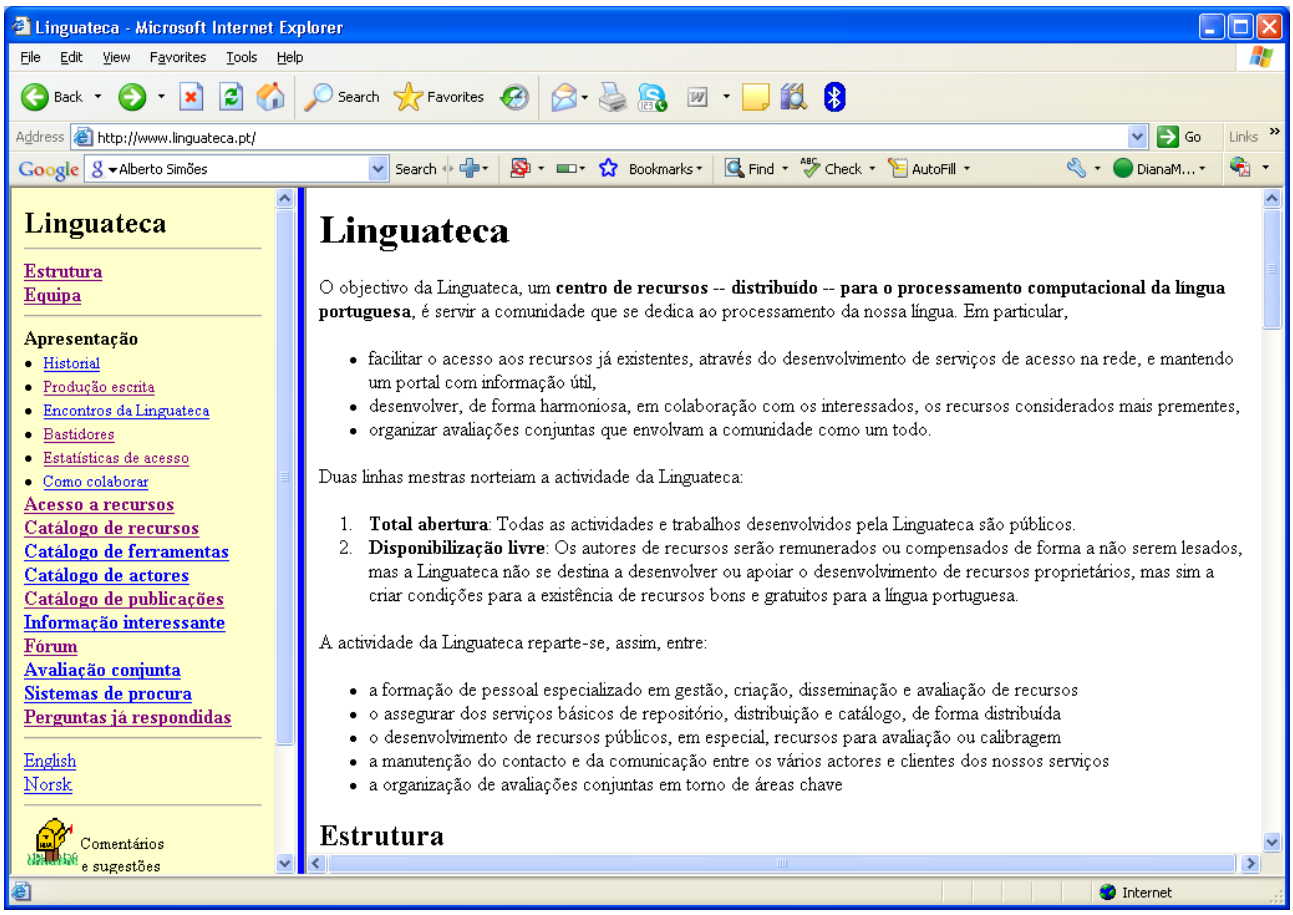

Figura 3: Ecrã da página de entrada da Linguateca

litários associados ao sítio da Linguateca (cujo desenho não é óbvio) pendem claramente para o lado da Linguateca e não da área em geral. Ou seja, as perguntas são exclusivamente sobre a Linguateca e os seus recursos, e os sistemas de procura têm como universo (ou base) todas as páginas apontadas pelo sítio da Linguateca mais as próprias páginas criadas por nós, o que significa, por definição, que incluem muito mais informação sobre a Linguateca do que sobre qualquer outro projecto na área.

Por um lado, isto pode compreender-se dado que é assim que funcionam todos os sistemas de busca locais (quem quer procurar de forma global e não local, usaria os motores gerais), mas, por outro lado, o objectivo de criar um sistema de busca na área, melhor do que os outros para esta área específica, porque informado por mais conhecimento, claramente falhou redondamente. Não por desígnio propositado, mas por o trabalho nessa ferramenta ter sido sempre preterido em relação a outros que pareciam mais urgentes ou que tinham utilizadores mais exigentes.

Provavelmente, este é um caso ovo-galinha clássico: nunca tivemos um sistema suficientemente bom para motivar utilizadores, donde estes nunca puxaram por nós, e por isso o sistema nunca foi desenvolvido como deveria.

Neste caso, a decisão e planeamento de quais as prioridades levou a que esse caminho ficasse atrofiado, muito embora a Linguateca até tenha aberto um pólo no grupo especializado nessa área em Portugal, o XLDB.

Voltando ao ponto de partida, é verdade que o sítio da Linguateca não se conseguiu impor como um catálogo actualizado, dinâmico e interessante para a área. Pelo contrário, a grande maioria dos nossos visitantes foram utilizadores dos recursos que criámos ou participantes nas actividades que organizámos.

Talvez também associado a esta questão, raríssimos foram os membros da comunidade que nos contactaram para incluirmos os seus recursos ou projectos no nosso sítio.

5.5.2 Falta de directivas

Embora tenhamos ganho muita experiência ao fazer e organizar avaliações conjuntas, medições de área e panorâmicas, não propagámos suficientemente (ou nada) como é que isso se deve fazer, como referido por Ferreira e Teixeira (2008).

Tal neste caso foi inocentemente motivado por imaginarmos que a Linguateca seria sempre o núcleo dessa organização, que grupos individuais

não se sentissem com motivação para levar aos ombros esse tipo de tarefa. Mas fica a chamada de atenção de que seria interessante tentar ensinar como fazer – refira-se que em Ferreira et al. (2009) já os mesmos autores demonstram a vantagem de o fazer no domínio da medicina.

5.5.3 Falta de ligação à comunidade empresarial

Outra crítica que nos foi feita, de formas variadas, foi que a Linguateca não olhou especialmente nem dedicou nenhuma vertente aos actores comerciais: assim, não só não nos preocupámos em ganhar dinheiro nem ajudar outros que connosco colaborassem a ganhá-lo, ou que quisessem colaborar connosco se nós os ajudássemos a ganhar dinheiro.

Embora eu não tenha a certeza de que concorde que isto deva ser visto como crítica – e de facto o testemunho de Braga e Dias (2008) pareça indicar que fomos, seja como for, úteis para algumas empresas, reconheço que é profundamente verdade.

Nós não dedicámos atenção diferente a nenhum tipo de actor e assumimos que a nossa actividade seria benéfica para todos por igual. Esta questão merece ser equacionada à luz destas críticas ou observações:

Seria aceitável ou (mais) produtivo se alguma actividade da Linguateca fosse dirigida (e mesmo paga) por actores comerciais, como aventado por Daniela Braga no encontro em Aveiro?

Seria natural transformar a Linguateca numa incubadora de empresas cujo objectivo seria rentabilizar e disseminar recursos públicos, como proposto por Anabela Barreiro no mesmo encontro?

Ficam as perguntas, e o repto de que esses modelos teriam de ser propostos e equacionados também por esses mesmos actores.

Aliás, e dada a (na minha opinião, triste) conversão progressiva das próprias universidades em máquinas de ganhar dinheiro, esta questão pode ser expandida a todos os modelos de colaboração com instituições no futuro.

O que não me parece fazer sentido, é propor que a Linguateca seja ela transformada numa actividade lucrativa.

5.6 Ferramentas em código aberto

Voltando a carregar na tecla "Casa de ferreiro, espeto de pau", o facto de o primeiro pólo da Linguateca em Portugal, o de Braga, ser especialista em código aberto e na disponibilização desse tipo de ferramentas não foi suficiente para conseguir que a Linguateca tivesse uma actividade consequente, profissional e de impacto profundo, quer na dita comunidade, quer em geral.

Com efeito, embora todo o código que tenhamos criado tenha vindo, melhor ou pior, a ser disponibilizado publicamente (o que não significa que tenha sido usado ou disseminado como deve ser), toda a cultura de desenvolvimento de código aberto não foi aproveitada, nem nós aproveitámos as possibilidades que teríamos de teste aos programas pela comunidade.

Por um lado, isso deveu-se ou deve-se à grande quantidade de linguagens de programação e ambientes usados, donde qualquer opção ou escolha nossa iria apenas satisfazer (ou melhor, apenas satisfez) um fragmento ou fracção da comunidade.[39]

Por outro lado, tivemos muitas vezes a impressão de que a maioria dos membros da comunidade preferiam obter programas a funcionar (e nesse caso como serviços na rede) do que estar a programar ou mexer em código de outrem. Os verdadeiros programadores, por outro lado, não abdicavam de programar tudo outra vez (de raiz) e estavam mais interessados em recursos ou ideias.

De qualquer maneira, temos de dar a mão à palmatória e confirmar que não conseguimos, nestes dez anos de actividade, produzir sistemas computacionais que fossem usados e manipulados por uma faixa grande de membros da nossa comunidade. Conseguimos isso em relação aos recursos, mas não a programas informáticos.

Embora também o NLP registry[40] seja um caso desses que parece não ter conseguido descolar[41], e que a maior parte dos programas de código aberto, mesmo no SourceForge, não têm sucesso (Feitelson, Heller e Schach, 2006), nós estamos claramente conscientes de que nos faltou uma estratégia nesse aspecto, assim como uma actividade de produção e manutenção dos sistemas já disponibilizados.[42] De facto, tal questão já tinha sido abordada criticamente em Santos (2000), mas não foi por isso resolvida.

Alguns exemplos de má prática:

O atomizador da Linguateca foi distribuído como um módulo do PLNbase pelo Alberto Simões, a cavalo noutro atomizador por ele desenvolvido (mas sem qualquer informação sobre as diferenças entre os dois). A primeira edição do atomizador e separador de frases foi publicada em 2004; desde essa altura e embora na Linguateca problemas pontuais e pequenas melhorias tenham

[39] A título anedótico, refira-se que, só dentro do âmbito da Linguateca, têm sido desenvolvidos e tornados públicos programas nas seguintes e diversas linguagens de programação: Perl, Java, PHP, C, R, Lisp, awk, Groovy e JavaScript.

[40] http://registry.dfki.de/

[41] Embora já na sua quarta versão, contém pouquíssimas entradas, e em muitas delas a informação sobre disponibilidade é simplesmente: "to negotiate".

[42] Tanto o catálogo de ferramentas, como o Jardim de Ferramentas, nunca tiveram de facto cobertura, publicidade e atenção suficientes para se tornarem eles próprios ferramentas úteis.

continuado a ser efectuadas, tal nunca (até agora) foi reflectido na versão pública.[43]

O Corpógrafo foi disponibilizado em código aberto antes de ser instalado em Barcelona,[44] mas o código ainda estava cheio de problemas e de questões não resolvidas, e só em fins de 2008 uma nova versão mais estável foi colocada ao dispor da comunidade. Este exemplo demonstra o que é bem sabido por todos os produtores comerciais: às vezes é preciso publicar ou pôr nas bancas um produto por razões que não são a de estar perfeito ou acabado. No nosso caso, foi para garantir que o produto seria tratado como código aberto pela instituição na qual foi instalado.

O código do Esfinge também foi disponibilizado desde 2006, veja-se Costa (2007), mas sem a garantia que as novas versões deste sistema, pioneiro para a língua portuguesa, estivessem logo acessíveis para a comunidade. Como só as pessoas que desenvolvem programas podem saber, não é trivial a documentação e manutenção de sistemas que evoluem ao longo de anos de trabalho, e existe sempre uma diferença entre uma versão estável e documentada e o programa do momento.

Finalmente, a questão da disponibilização de sistemas complexos ainda provoca mais dificuldade devido à questão das dependências: não faz sentido começar a fazer tudo do nada, mas, se se inclui outros sistemas, como seria natural e boa prática, obriga-se o utilizador incauto a instalar e ter de levar em conta muitos outros programas desenvolvidos por terceiros e que podem eles próprios ser difíceis de instalar ou compreender.

5.7 Documentação – a sempre vilipendiada

Há duas leis na informática: a de que a documentação é essencial, e a de que a documentação nunca está actualizada. Todos os projectos lutam com estas duas leis, e embora no caso da Linguateca tenhamos feito um esforço não irrisório de boa doumentação, não conseguimos também escapar à segunda lei, de que ainda falta documentar ou melhorar muita coisa.

Ao contrário do que certas pessoas pregam, de que um programa ou sistema bom ou bem desenhado não precisa de explicação ou documentação, tal parece-me completamente errado no caso da área do processamento de uma língua. Não vou pois argumentar em geral, mas apenas no domínio em que trabalhamos.

Dando alguns exemplos concretos:

- qual a utilidade de saber quantos substantivos ou adjectivos há num texto, sem saber quais os critérios de classificação de uma e outra categoria?
- qual a utilidade de saber quais as palavras mais frequentes, ou a frequência de um conjunto de palavras, sem se saber qual a base (os textos) usada para essas contagens?
- que vantagem tem um sistema que anota um texto, sem que se saiba os critérios de anotação usados?

Ou: como é que se pode avaliar um dado sistema se não se consegue interpretar a sua saída? Como é que se pode usar um sistema para fazer uma coisa quando foi desenhado para outra?

Em todos os casos de trabalho sério, é preciso saber como é que cada tarefa ínfima é feita – ou ter a possibilidade de o saber. Sem isso, estamos no reino da "banha da cobra", e não estamos a criar recursos ou ferramentas que possam contribuir para o progresso e que possam ser melhorados por outros. Estamos apenas a tentar vender, no sentido de convencer a usar, um produto de forma irresponsável.

Este aspecto da documentação e da explicação de como é que os recursos foram criados, e quais os pressupostos envolvidos na sua criação, é uma das tónicas mais importantes postas pela Linguateca no seu trabalho.

Outra questão – menos crítica – é a remoção de assuntos ou páginas claramente desactualizadas ou irrelevantes, que tendem a ficar perdidas ou penduradas num sítio da rede em vez de activamente limpas ou reescritas pelos gestores do sítio. Embora isto faça parte do manual dos gestores de sítios, é preciso reconhecer ou relembrar que as principais capacidades da Linguateca não são a de gestão profissional de sítios. Apenas muito recentemente, há menos de um ano, passámos a gerir uma parte (ínfima) das nossas actividades em wiki, como se pode ver em relação à página do GikiCLEF. Tal deveu-se, mais uma vez, a não haver pessoal com apetência especial para manutenção de sítios e ao facto de termos já uma quantidade de programas e rotinas desenhadas para gerir o sítio da Linguateca, e que reconvertê-las levaria a muito trabalho – que seria afinal só cosmético.

Assim, embora a documentação e a apresentação sejam de certa forma acessórias ao verdadeiro trabalho da Linguateca, são requisitos necessários para que este seja compreendido e usado. Sistemas ou serviços sem documentação, são completamente inúteis – ou até perigosos, se induzirem

[43] A reforçar o já dito anteriormente sobre as linguagens de programação, uma total reescrita do mesmo atomizador noutra linguagem foi recentemente disponibilizada por Nuno Cardoso no âmbito do seu sistema REMBRANDT (Cardoso, 2008c).

[44] No âmbito da colaboração entre o CLUP/Linguateca e o grupo de Teresa Cabré no Institut Universitari de Lingüística Aplicada (IULA) na Universitat Pompeu Fabra.

as pessoas em erro.

Mas sistemas e serviços que devido à sua má apresentação assustam ou repelem os utilizadores a quem foram destinados também constituem um entrave sério ao impacto da Linguateca e à nossa possibilidade de sermos úteis à comunidade.

5.8 A usabilidade e preocupação com os utilizadores

De facto, uma outra área que é preciso mencionar, é a usabilidade, ou seja, a preocupação da Linguateca com os utilizadores dos vários programas que desenvolvemos, avaliamos ou estudamos. Pese embora a nossa consciencialização sobre o assunto, e uma tentativa de actuação variada, o cômputo geral parece mais negativo do que positivo.

Esta preocupação pode apreciar-se em vários ramos diferentes da nossa intervenção na área do processamento da língua:

Por um lado, refira-se o estudo sério de necessidades de informação como preliminar para o desenvolvimento posterior do sistema de recolha de informaçao na rede de Rachel Aires (Aires e Aluísio, 2003), que aliás fez girar toda a problemática da sua tese à volta da formalização e detecção das necessidades do utilizador, e efectuou testes com utilizadores para avaliar o sistema implementado.

Por outro, tivemos sempre uma atitude muito crítica em relação à forma como algumas tarefas foram definidas no CLEF, pondo-nos no lugar de utilizadores de língua portuguesa, ou de simples pessoas interessadas em recolha de informação cruzada (Santos e Rocha, 2005; Santos e Cardoso, 2005). Em muitas ocasiões, fomos de certa forma os primeiros a gritar que "o rei vai nu": muitas das hipóteses tomadas como óbvias num ambiente anglofalante caem pela base ao considerar outras línguas, no nosso caso o português.

Como já mencionado, fomos dos primeiros a nível internacional a levar a cabo, e a publicar, dados sobre utilizadores de um serviço de corpos, o COMPARA (Santos e Frankenberg-Garcia, 2007), em que explicitamente aplicamos métodos de investigação não-obstrusiva da actividade dos utilizadores aos diários de interacção com o serviço.

Fomos também dos primeiros a executar estudos dos diários de procura na rede com base no instantâneo da rede portuguesa WPT03 para efeitos de processamento da língua ou recolha de informação (Seco e Cardoso, 2006).

Finalmente, a um nível completamente diferente, implementámos um serviço cooperativo de resposta aos utilizadores de forma a dar sempre resposta às mais variadas questões, como mencionado na secção anterior.

Contudo, a aparência dos nossos serviços e informação na rede foi sempre o nosso calcanhar de Aquiles e, nas palavras críticas de um dos leitores do presente artigo:

> É uma imagem que me transporta para meados dos anos 90. (...) qualquer utilizador banal vai pensar que o site não é actualizado há anos e que não vai encontrar lá nada de útil. Transmite a ideia de site criado por amadores, sem conhecimentos de informática.

Numa altura em que todas as empresas, pelo menos as associadas a meios de comunicação social ou editorial, aplicam rotineiramente análise de diários e de comportamento de utilizadores para melhorar a sua presença na rede, a Linguateca, embora possivelmente à frente na comunidade científica do processamento da língua, está muito atrás da realidade da vida de todos os dias.

5.9 Publicação em nome da Linguateca

Embora a Linguateca possa apregoar um grande número de publicações e apresentações produzidos ao longo destes dez ou onze anos – trezentas a quatrocentas, não podemos infelizmente garantir ou confirmar que todos os textos publicados com a chancela da Linguateca tenham sido verificados em termos de qualidade ou mesmo de oportunidade.

A existência de cerca de trinta colaboradores ao longo do tempo e o facto de as publicações não estarem prontas na maior parte das vezes a tempo suficiente antes da data final de entrega levou a uma publicação muito descentralizada e que não usufruiu, na maior parte dos casos, das vantagens que poderia colher ao ser redigida no seio de um equipa de peritos.

Isso, aliás, é claramente patente na ausência, na maior parte dos artigos, de agradecimentos a revisão cruzada de outros elementos da Linguateca. Não dizendo que isto é um problema específico da nossa equipa, falhou claramente, na maior parte dos casos, também entre nós a possibilidade de retorno e de discussão científica séria antes da publicação.

Idealmente, deveríamos ter definido normas mais concretas tanto quanto à divulgação da Linguateca em geral como ao posicionamento do trabalho relatado no plano geral da nossa actividade, assim como deveríamos ter estipulado um certo conjunto de normas de qualidade, empíricas, a que os artigos da Linguateca como Linguateca deviam obedecer, e que em alguns casos teriam levado a uma reescrita ou à não publicação do artigo como trabalho realizado no âmbito da Linguateca. Se viermos a continuar como instituição virtual, parece-me que isto tem de ser decididamente contemplado no futuro, até porque teria sido uma forma relati-

vamente fácil de obter maior impacto.

Que é possível empenhar a equipa – e mesmo elementos de fora da Linguaetca mas que possam rever-se como pertencendo ao círculo da mesma – foi patente em relação ao presente texto, o qual foi extraordinariamente melhorado devido ao excelente retorno e problematização de várias afirmações e opiniões patentes em versões anteriores, por mais de uma dezena de leitores interessados.

6 A saúde do processamento computacional do português

Embora este artigo seja sobre a Linguateca, não posso deixar de chamar aqui a atenção sobre outras vitórias nesta área durante o período coberto por esta reflexão, completamente independentes da nossa acção. Não gostava de forma nenhuma de parecer estar a afirmar que, sem nós, nada teria acontecido, ou que, excepto nós, ninguém fez nada.

Assim, gostava de salientar – sem quaisquer pretensões de exaustividade, visto que tal assunto poderia e deveria constituir um artigo novo – alguns acontecimentos ou sistemas que me parece fazerem a diferença, ou seja, serem vitórias incontornáveis do português no campo internacional:

- o primeiro detector automático de metáforas foi desenvolvido para o português – e depois aplicado ao inglês – por Tony Berber Sardinha (Berber Sardinha, 2006; Berber Sardinha, 2007);
- o primeiro sistema automático para produção de livros auditivos foi criado por uma parceria entre o INESC e a FCUL (Serralheiro et al., 2003);
- o primeiro serviço automático com classificação semântica foi feito no VISL para o português (Bick, 2006; Bick, 2007)[45];
- o primeiro motor de procura sobre a rede completa de um país foi efectuado pela equipa do tumba! (Gomes e Silva, 2005);
- a primeira legendagem automática de telejornais para deficientes auditivos foi realizada pelo projecto Tecnovoz (Meinedo, Viveiros e Neto, 2008);
- a primeira geração de fala para fórmulas matemáticas ou equações foi descrita em Rolo e Serralheiro (2008).

[45] É preciso notar que embora Eckhard Bick tenha uma relação estreita com a Linguateca, a grande maioria dos rtrabalhos efectuados pelo projecto VISL são completamente independentes desta. O que também se aplica ao grupo do XLDB ou outros que sejam mencionados nesta secção.

Mesmo quando não estamos a falar de primeiros para qualquer língua, não queremos deixar de chamar a atenção, que, para o português, houve naturalmente muitíssimos "primeiros" sem qualquer relação com a Linguateca.

Por exemplo, os três seguintes sistemas ou recursos nasceram no NILC:

- o primeiro sistema de sumarização automática para o português (Pardo e Rino, 2002);
- a primeira ontologia lexical para o português inspirada pelo método da WordNet (Oliveira, Dias da Silva e Moraes, 2002);
- o primeiro detector da estrutura retórica de um texto para o português (Pardo, Nunes e Rino, 2004).

E outros primeiros foram:

- o primeiro sistema de RAP em português baseado em análise sintáctica, pelo VISL (Bick, 2003);
- o primeiro sistema completo de síntese de base articulatória suportada em estudos de produção para o português, pelo IETA em Aveiro (Oliveira, 2009);
- o primeiro sistema de desenvolvimento de ontologias a partir de texto pela PUC-RS (Gasperin, 2001);
- o primeiro modelo cognitivo quantitativo para o estudo da evolução diacrónica de variedades do português (Silva, 2008a).

Tal é sinal evidente de que o processamento do português tem boas pernas para andar. Penso que – de preferência com a colaboração de todos – poderemos ir longe na investigação e desenvolvimento de sistemas computacionais que lidem perfeitamente com a nossa língua.

7 Comentários finais

Neste artigo, comecei por comparar as intenções iniciais e o ponto de situação efectuado no começo da actividade da Linguateca, como um exercício salutar de avaliação, dez anos passados. Apresentei brevemente a história da Linguateca, depois salientei sucintamente as actividades ou áreas de intervenção em que penso que a Linguateca foi útil para a comunidade do processamento do português e nem só, passando a indicar os problemas ou áreas em relação aos quais a Linguateca não conseguiu, na minha opinião, dar um contributo suficientemente positivo.

Tentei mostrar que ao longo da nossa história muito de bom aconteceu, apresentando alguns casos de maturidade e de inovação na área. Também

considero, contudo, que muito mais podia ter sido feito se tivesse havido confiança na Linguateca e um espírito de colaboração entre os vários grupos ou instituições dedicados à área, especialmente em Portugal. Espírito esse que foi apanágio de muito dos nossos colegas brasileiros, que cooperaram, produziram recursos para o repositório, e aproveitaram (como nós queríamos) o nosso trabalho, e a quem estou particularmente grata por isso.

Se pudesse começar de novo, e mais uma vez esta é uma visão muito pessoal, continuaria a organizar avaliações conjuntas e a criar recursos de avaliação em conjunto com membros da comunidade, mas não tentaria catalogar a área ou observá-la, tentando fixá-la num sítio megalómano. Pelo contrário, tentaria que todos discutissem e comunicassem através de listas de discussão e da troca de ideias e, claro, da participação em avaliações conjuntas.

Assim como temos um serviço de resposta a todas as perguntas que nos fazem (mas que são limitadas e muitas vezes fora do contexto da própria Linguateca), tentaria fazer com que essas perguntas fossem feitas e respondidas num verdadeiro fórum de todos os interessados na área (como acontece por exemplo na lista `corpora`), permitindo a interacção, o conhecimento dos intervenientes, e uma resposta cooperativa que ajuda a quem perguntou mas também aos outros que estão a ouvir porque fazem parte da comunidade.

Tentaria também oferecer a Linguateca como um serviço de avaliação no sentido de podermos ajudar a criar materiais de teste ou mesmo métricas para avaliar trabalhos ou sistemas de empresas ou académicos, devido à nossa experiência no assunto.

Finalmente, se fosse a continuação da Linguateca que estava em jogo, e nos fossem concedidos mais dez anos, seria essencial focar-nos em projectos com impacto nacional ou internacional (em língua portuguesa, claro), tal como o Museu da Pessoa, a procura inteligente nas obras da(s) Biblioteca(s) Nacional(is), a procura na rede, o arquivo da rede portuguesa e brasileira, e sistemas de tradução automática com respeito pelo português, não descurando, também, toda a parte cultural e multimodal associada à procura em imagens, vídeo e sons, e em meios mistos.

É minha convicção de que uma Linguateca futura teria de ter uma componente prática muito maior envolvendo empresas e instituições, e o seu fito deveria ser aplicar a tecnologia existente à realidade de todos os dias.

Não faz sentido a continuação da Linguateca como é agora, apenas com parceiros académicos e com impacto na comunidade científica: a Linguateca para merecer sobreviver e poder continuar a ser útil, terá de se "praticalizar", ou seja, tomar em mãos aspectos e projectos claramente práticos.

Agradecimentos

Este artigo foi escrito no âmbito da Linguateca, contrato número 339/1.3/C/NAC, financiado pelo governo português e pela União Europeia.

A existência da Linguateca deve-se, em primeiro lugar, ao interesse do então ministro da Ciência e da Tecnologia, José Mariano Gago, pela questão da língua, que levou à inclusão deste assunto no Livro Verde e depois no Livro Branco, e, em segundo lugar, ao apoio constante, institucional e pessoal, do presidente da FCCN[46], Pedro Veiga.

Agradeço a todos os membros da Linguateca, a todas as pessoas que colaboraram com a Linguateca, a todos os que contribuíram, com as suas perguntas, pedidos ou sugestões, para a melhoria do nosso projecto, e finalmente a todos os que comentaram, criticaram e enriqueceram o presente texto.

Referências

Afonso, Susana. 2003. Clara e sucintamente: um estudo em corpus sobre a coordenação de advérbios em -mente. Em Amália Mendes e Tiago Freitas, editores, *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002)*, pp. 27–36, Lisboa, 2-4 de Outubro, 2003. APL.

Afonso, Susana. 2004. Estudo dos argumentos verbais e ambiguidade dos sintagmas preposicionais através do Águia. Relatório técnico, Linguateca, 21 de Abril, 2004. `http://www.linguateca.pt/documentos/ArgumentosambiguidadeAfonso2004.pdf`.

Afonso, Susana, Eckhard Bick, Renato Haber, e Diana Santos. 2001. Floresta sintá(c)tica: um treebank para o português. Em Anabela Gonçalves e Clara Nunes Correia, editores, *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001)*, pp. 533–545, Lisboa, Portugal, 2-4 de Outubro, 2001. APL.

Aires, Rachel e Diana Santos. 2002. Measuring the Web in Portuguese. Em Brian Matthews, Bob Hopgood, e Michael Wilson, editores, *Euroweb 2002 conference*. pp. 198–199, 17-18 Dezembro, 2002.

Aires, Rachel Virgínia Xavier. 2005. *Uso de marcadores estilísticos para a busca na Web em por-*

[46] A FCCN é a instituição portuguesa que, em termos jurídicos, é "executora" do projecto Linguateca desde 2000.

tuguês. Tese de doutoramento, ICMC - USP - São Carlos, Agosto, 2005.

Aires, Rachel Virgínia Xavier e Sandra Maria Aluísio. 2003. Como incrementar a qualidade das máquinas de busca: da análise de logs à interação em português. *Revista Ciência da Informação*, 32(1):5–16.

Almeida, José João, editor. 2003. *Corpora Paralelos, Aplicações e Algoritmos Associados (CP3A)*. Universidade do Minho, Braga.

Almeida, José João e Alberto Simões. 2007. XML::TMX - Processamento de Memórias de Tradução de Grandes Dimensões. Em José Carlos Ramalho, João Correia Lopes, e Luís Carriço, editores, *XML: Aplicações e Tecnologias Associadas (XATA2007)*, pp. 83–93. Universidade do Minho, 15-16 de Fevereiro, 2007.

Almeida, José João, Alberto Manuel Simões, e José Alves Castro. 2002. Grabbing parallel corpora from the web. *Sociedade Española para el Procesamiento del Lenguaje Natural*, 29:13–20.

Aluisio, Sandra, Gisele Montilha Pinheiro, Aline M. P. Manfrin, Leandro H. M. de Oliveira, Luiz C. Genoves Jr., e Stella E. O. Tagnin. 2004. The Lácio-Web: Corpora and tools to advance Brazilian Portuguese language investigations and computational linguistic tools. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, pp. 1779–1782, 26-28 de Maio, 2004.

Aluísio, Sandra Maria, Leandro H.M. de Oliveira, e Gisele Montilha Pinheiro. 2004. Os tipos de anotações, a codificação, e as interfaces do Projeto Lácio-Web: Quão longe estamos dos padrões internacionais para córpus? Em *II Anais do TIL - Workshop de Tecnologia da Informação e Linguagem Humana*, pp. 1–10, 5 a 6 de Agosto, 2004.

Amaral, Carlos, Helena Figueira, Afonso Mendes, Pedro Mendes, Cláudia Pinto, e Tiago Veiga. 2008. Adaptação do sistema de reconhecimento de entidades mencionadas da Priberam ao HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Aranha, Christian Nunes. 2007. O Cortex e a sua participação no HAREM. Em Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pp. 113–122.

Bacelar do Nascimento, Maria Fernanda, Amália Mendes, e Luísa Pereira. 2004. Providing online access to portuguese language resources: corpora & lexicons. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, pp. 1825–1828, 26-28 de Maio, 2004.

Barreiro, Anabela. 2008. ParaMT: a Paraphraser for Machine Translation. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, e Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume Vol. 5190. Springer Verlag, pp. 202–211, 8-10 de Setembro, 2008.

Barreiro, Anabela e Susana Afonso. 2007. Construção da lista dourada para as primeiras Morfolimpíadas do português. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, pp. 107–118.

Barreiro, Anabela e Elisabete Ranchhod. 2005. Machine Translation Challenges for Portuguese. *Linguisticæ Investigationes*, 28(1):3–18.

Berber Sardinha, Tony. 2006. An online program for tagging metaphors in corpora. Em S. Zyngier, V. Viana, e A. M. Spallanzani, editores, *Linguagens e Tecnologias: Estudos Empíricos*, pp. 165–182, Rio de Janeiro, Brasil. Editora da UFRJ.

Berber Sardinha, Tony. 2007. *Metáfora*. Parabola, São Paulo, Brasil.

Berber Sardinha, Tony, J. L. Moreira Filho, e E. Alambert. 2008. O corpus brasileiro. Comunicação ao VII Encontro de Lingüística de Corpus, 2008, UNESP, São José do Rio Preto, SP, Brasil.

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

Bick, Eckhard. 2003. A Constraint Grammar Based Question-Answering System for Portuguese. Em Fernando Moura Pires e Salvador Abreu, editores, *Progress in Artificial Intelligence: 11th Portuguese Conference on Artificial Intelligence, EPIA 2003. Beja, Portugal, December 2003, Proceedings*, pp. 414–418, Berlin/Heidelberg. Springer.

Bick, Eckhard. 2006. Noun sense tagging: Semantic prototype annotation of a portuguese treebank. Em Jan Hajic e Joakim Nivre, editores, *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT 2006)*, 1-2 de Dezembro, 2006.

Bick, Eckhard. 2007. Automatic semantic role annotation for portuguese. Em *TIL, V Workshop em Tecnologia da Informação e da Linguagem Humana*, pp. 1715–1719, 30 de Junho a 6 de Julho, 2007.

Bick, Eckhard, Diana Santos, Susana Afonso, e Rachel Marchi. 2007. Floresta Sintá (c)tica: Ficção ou realidade? Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, pp. 291–300.

Braga, Daniela e Miguel Sales Dias. 2008. Os recursos da Linguateca ao serviço do desenvolvimento da tecnologia de fala na Microsoft. Em Luís Costa, Diana Santos, e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, pp. 29–33.

Cabral, Luís Miguel. 2007. SUPeRB - Sistema Uniformizado de Pesquisa de Referências Bibliográficas. Tese de Mestrado, Faculdade de Engenharia da Universidade do Porto, Porto, Março, 2007.

Cabral, Luís Miguel, Luís Fernando Costa, e Diana Santos. 2007. Esfinge at CLEF 2007: First steps in a multiple question and multiple answer approach. Em Alessandro Nardi e Carol Peters, editores, *Working Notes for the CLEF 2007 Workshop (CLEF 2007)*, pp. s/pp, 19-21 de Setembro, 2007.

Cabral, Luís Miguel, Diana Santos, e Luís Fernando Costa. 2008. SUPeRB - Gerindo referências de autores de língua portuguesa. Em *VI Workshop Information and Human Language Technology (TIL'08)*, 28-29 de Outubro, 2008.

Calado, Pável. 1999. The WBR-99 Collection: Description of the WBR-99 Web collection data-structures and file formats. Relatório técnico, LATIN - Laboratório para o Tratamento de Informação, Departamento de Computação, Universidade Federal de Minas Gerais. `http://www.linguateca.pt/Repositorio/WBR-99/wbr99.pdf`.

Cardoso, Nuno. 2008a. Apêndice H: SAHARA - Serviço de Avaliação HAREM Automático. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Cardoso, Nuno. 2008b. Novos rumos para a recuperação de informação geográfica em português. Em Luís Costa, Diana Santos, e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*. Linguateca, pp. 71–85.

Cardoso, Nuno. 2008c. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Cardoso, Nuno, Bruno Martins, Daniel Gomes, e Mário J. Silva. 2007. WPT 03: a primeira colecção pública proveniente de uma recolha da web portuguesa. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, pp. 279–288.

Cardoso, Nuno e Diana Santos. 2007. Directivas para a identificação e classificação semântica na colecção dourada do HAREM. Em Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pp. 211–238.

Carvalho, Paula, Hugo Gonçalo Oliveira, Diana Santos, Cláudia Freitas, e Cristina Mota. 2008. Segundo HAREM: Modelo geral, novidades e avaliação. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Chaves, Marcirio, Catarina Rodrigues, e Mário J. Silva. 2007. Data Model for Geographic Ontologies Generation. Em José Carlos Ramalho, João Correia Lopes, e Luís Carriço, editores, *XML: Aplicações e Tecnologias Associadas (XATA2007)*, pp. 47–58. Universidade do Minho, 15-16 de Fevereiro, 2007.

Chaves, Marcirio Silveira. 2008. *Uma Metodologia para Construção de Geo-Ontologias*. Tese de doutoramento, Faculdade de Ciências, Universidade de Lisboa, Dezembro, 2008.

Chinchor, Nancy e P. Robinson. 1998. MUC-7 Named Entity Task Definition (version 3.5). Em *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, Virginia, EUA.

Chubin, Daryl E. e Edward J. Hackett. 1990. *Peerless Science: Peer Review and U.S. Science*

Policy. State University of New York Press, Nova Iorque, EUA.

Costa, Luís. 2005. Esfinge - Resposta a perguntas usando a Rede. Em José María Gutiérrez, Flavia Maria Santoro, e Pedro Isaías, editores, *Proceedings da conferência IADIS Ibero-Americana WWW/Internet 2005*, pp. 616–619. IADIS Press, 18-19 de Outubro, 2005.

Costa, Luís. 2006. Esfinge - A Question Answering System in the Web using the Web. Em *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pp. 127–130, 3-7 de Abril, 2006.

Costa, Luís. 2007. Question answering beyond CLEF document collections. Em Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, e Maximilian Stempfhuber, editores, *Evaluation of Multilingual and Multimodal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers*, volume 4730 of *Lecture Notes in Computer Science*. Springer, Berlin / Heidelberg, pp. 405–414.

Costa, Luís. 2008. Resumo da actividade da Linguateca de 16 de Dezembro de 2006 a 31 de Dezembro de 2008. Relatório técnico, Linguateca, Dezembro, 2008. Com a colaboração (por ordem alfabética) de Ana Frankenberg-Garcia, Anabela Barreiro, Cláudia Freitas, Cristina Mota, David Cruz, Diana Santos, Hugo Oliveira, Luís Cabral, Nuno Cardoso, Paula Carvalho Paulo Rocha, Sérgio Matos, `http://www.linguateca.pt/documentos/RelatorioLinguateca20072008.pdf`.

Costa, Luís e Luís Miguel Cabral. 2008. Medindo a Linguateca, 11 de Setembro, 2008. `http://www.linguateca.pt/Linguateca10anos/Apresentacoes/AprCostaCabralL10.pdf`.

Costa, Luís, Diana Santos, e Nuno Cardoso, editores. 2008. *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca.

Davies, Mark e Ana Maria Raposo Preto-Bay. 2008. The Corpus do Português and the Routledge frequency dictionary of Portuguese: New tools for learners and teachers. Em *TaLC 8 Lisbon: Proceedings of 8th Teaching and Language Corpora Conference (3-6 July 2008)*. Associação de Estudos e de Investigação Científica do ISLA - Lisboa, pp. 96–99.

Feitelson, Dror G., Gillian Z. Heller, e Stephen R. Schach. 2006. An empirically-based criterion for determining the success of an open-source project. Em *Australian Software Engineering Conference*, pp. 363–368, Abril, 2006.

Fernandes, Eraldo R., Ruy L. Milidiu, e Cicero N. Santos. 2009. Portuguese language processing service. Em *18th International World Wide Web Conference*, 20-24 de Abril. 2009.

Ferreira, Liliana, Cesar Telmo Oliveira, António Teixeira, e João Paulo Silva Cunha. 2009. Extracção de informação de relatórios médicos. *Linguamática*, 1, Maio, 2009.

Ferreira, Liliana e António Teixeira. 2008. Linguateca e Processamento de Linguagem Natural na Área da Saúde: Alguns Comentários e Sugestões. Em Luís Costa, Diana Santos, e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, pp. 43–48, 11 de Setembro, 2008.

Forner, Pamela, Anselmo Peñas, Iñaki Alegria, Corina Forascu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, Richard Sutcliffe, e Erik Tjong Kim Sang. 2009. Overview of the CLEF 2008 Multilingual Question Answering Track. Em Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, e Viviane Petras, editores, *Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Springer.

Frankenberg-Garcia, Ana e Diana Santos. 2002. COMPARA, um corpus paralelo de português e de inglês na Web. *Cadernos de Tradução*, IX(1):61–79.

Freitas, Cláudia. 2008. A Floresta Sintáctica no Ensino de Português, 3 de Julho, 2008. `http://www.linguateca.pt/documentos/FreitasWorkshopTaLC2008.pdf`.

Freitas, Cláudia e Susana Afonso. 2008. Bíblia Florestal: Um manual lingüístico da Floresta Sintá (c)tica. `http://linguateca.dei.uc.pt/Floresta/BibliaFlorestal/`.

Freitas, Cláudia, Paulo Rocha, e Eckhard Bick. 2008a. Um mundo novo na Floresta Sintá (c)tica - o treebank para Português. *Calidoscópio - Revista de Pós Graduação em Lingüística Aplicada da Unisinos, Rio Grande do Sul*, 6(3), Set / Dezembro, 2008.

Freitas, Cláudia, Paulo Rocha, e Eckhard Bick. 2008b. Um mundo novo na Floresta Sintá

(c)tica - o treebank para Português. *Calidoscópio - Revista de Pós Graduação em Lingüística Aplicada da Unisinos, Rio Grande do Sul*, 6(3), Set / Dezembro, 2008.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, e Cristina Mota. 2008. Relações semânticas do ReRelEM: além das entidades no Segundo HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 31 de Dezembro, 2008.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, e Cristina Mota. 2009. Relation detection between named entities: report of a shared task. Em *Proceedings of Semantic Evaluations Workshop*, 4 de Junho, 2009.

Gasperin, Caroline Varaschin. 2001. Extração automática de relações semânticas a partir de relações sintáticas. Tese de Mestrado, Faculdade de Informática da Pontifícia Universidade Católica do Rio Grande do Sul.

Gomes, Daniel e Mário J. Silva. 2005. Characterizing a National Community Web. *ACM Transactions on Internet Technology*, 5(3):508–531, Agosto, 2005.

Gomes, Paulo. 2008. Linguateca: Polo de Coimbra - Plantando Florestas e Criando Papel, 11 de Setembro, 2008. `http://www.linguateca.pt/Linguateca10anos/Apresentacoes/AprGomesL10.pdf`.

Gomes de Matos, Francisco. 1992. O cientista de língua portuguesa e seus direitos linguísticos. *Revista Internacional de Língua Portuguesa*, 7:79–81.

Gonçalo Oliveira, Hugo, Cristina Mota, Cláudia Freitas, Diana Santos, e Paula Carvalho. 2008a. Avaliação à medida no Segundo HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 31 de Dezembro, 2008.

Gonçalo Oliveira, Hugo, Diana Santos, Paulo Gomes, e Nuno Seco. 2008b. PAPEL: a dictionary-based lexical ontology for Portuguese. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, e Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, pp. 31–40. Springer Verlag, 8-10 de Setembro, 2008.

Haber, Renato Ribeiro. 2001. Pica-pau: Um protótipo de ferramenta para visualização e edição de árvores sintáticas. Texto produzido no âmbito da Floresta Sintá (c)tica, `http://www.linguateca.pt/treebank/Picapau.html`.

Hausser, Roland, editor. 1996. *Linguistische Verifikation: Dokumentation zur Ersten Morpholympics 1994*. Max Niemeyer Verlag.

Inácio, Susana e Diana Santos. 2006. Syntactical Annotation of COMPARA: Workflow and First Results. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira, e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006*, volume LNAI 3960, pp. 256–259, Berlin/Heidelberg, 13-17 de Maio, 2006. Springer.

Inácio, Susana e Diana Santos. 2008. Documentação da anotação morfossintáctica da parte portuguesa do COMPARA, Dezembro, 2008. Primeira versão: 9 de Dezembro de 2005, `http://www.linguateca.pt/COMPARA/DocAnotacaoPortCOMPARA.pdf`.

Inácio, Susana, Diana Santos, e Rosário Silva. 2008. COMPARAndo cores em português e inglês. Em Sónia Frota e Ana Lúcia Santos, editores, *Artigos seleccionados do XXIII Encontro da Associação Portuguesa de Linguística (APL)*, pp. 271–286, 1-3 de Outubro de 2007, 2008.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, e David Tugwell. 2005. The Sketch Engine. Em *Proc. Euralex*. pp. 105–116, Julho, 2005.

Lai, Catherine e Steven Bird. 2004. Querying and updating treebanks: A critical survey and requirements analysis. Em *In Proceedings of the Australasian Language Technology Workshop*, pp. 139–146.

Maia, Belinda. 2003. Constructing comparable and parallel corpora for terminology extraction - work in progress. Em Dawn Archer, Paul Rayson, Andrew Wilson, e Tony McEnery, editores, *Proceedings of the Corpus Linguistics 2003 conference (CL2003)*, 28-31 de Março. 2003.

Maia, Belinda. 2008a. Alice no País das Maravilhas ou as aventuras e desventuras de uma linguista no mundo do PLN, 11 de Setembro, 2008. `http://www.linguateca.pt/Linguateca10anos/Apresentacoes/AprMaiaL10.pdf`.

Maia, Belinda. 2008b. Corpógrafo V4 - Tools for Educating Translators. Em Elia Yuste Rodrigo,

editor, *Topics in Language Resources for Translation and Localisation*. John Benjamins Pub. Co, Amsterdam/Philadelphia, pp. 57–70, Novembro, 2008.

Maia, Belinda e Anabela Barreiro. 2007. Uma experiência de recolha de exemplos classificados de tradução automática de inglês para português. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, pp. 205–216, 20 de Março, 2007.

Maia, Belinda e Sérgio Matos. 2008. Corpógrafo V4 - Tools for Researchers and Teachers using Comparable Corpora. Em Pierre Zweigenbaum, Éric Gaussier, e Pascale Fung, editores, *LREC 2008 Workshop on Comparable Corpora (LREC 2008)*. European Language Resources Association (ELRA), pp. 79–82, 31 de Maio, 2008.

Maia, Belinda, Luís Sarmento, e Diana Santos. 2005. Introduzindo o Corpógrafo - um conjunto de ferramentas para criar corpora especializados e comparáveis e bases de dados teminológicas. *Terminómetro*, 7:61–62. Número especial - A terminologia em Portugal e nos países de língua portuguesa em África.

Meinedo, Hugo, Márcio Viveiros, e João Paulo da Silva Neto. 2008. Evaluation of a live broadcast news subtitling system for Portuguese. Em *Interspeech 2008*. ISCA, Setembro, 2008.

Mota, Cristina e Pedro Moura. 2003. ANELL: A Web System for Portuguese Corpora Annotation. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003*, pp. 184–188, Berlin/Heidelberg. Springer Verlag.

Mota, Cristina e Diana Santos, editores. 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

NIST e ACE. 2007. Automatic Content Extraction 2008 Evaluation Plan (ACE08) – Assessment of Detection and Recognition of Entities and Relations within and across Documents. Relatório técnico, NIST. `http://www.nist.gov/speech/tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf`.

Nunes, Maria das Graças Volpe. 2008. Relato sobre a parceria Linguateca-NILC, 11 de Setembro, 2008. `http://www.linguateca.pt/Linguateca10anos/Apresentacoes/AprNunesL10.pdf`.

Oliveira, Catarina Alexandra Monteiro de. 2009. *Do Grafema ao Gesto: Contributos Linguísticos para um Sistema de Síntese de Base Articulatória*. Tese de doutoramento, Universidade de Aveiro.

Oliveira, Débora, Luís Sarmento, Belinda Maia, e Diana Santos. 2005. Corpus analysis for indexing: when corpus-based terminology makes a difference. Em Pernilla Danielsson e Martijn Wagenmakers, editores, *Proceedings from the Corpus Linguistics 2005 Conference Series*, volume 1, 14-17 de Julho. 2005.

Oliveira, Mirna, Bento C. Dias da Silva, e Helio Moraes. 2002. Groundwork for the Development of the Brazilian Portuguese Wordnet. Em Nuno Mamede e Elisabete Ranchhod, editores, *Advances in Natural Language Processing: Third International Conference, Proceedings (PorTAL 2002)*, Lecture Notes in Artificial Intelligence, pp. 189–196, Berlin/Heidelberg, 23-26 de Junho, 2002. Springer.

Orăsan, Constantin, Dan Cristea, Ruslan Mitkov, e Antonio Branco. 2008. Anaphora resolution exercise: An overview. Em *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marraqueche, Marrocos, 28 - 30 de Maio, 2008.

Pardo, Thiago A. S., Maria das Graças Volpe Nunes, e Lúcia H. M. Rino. 2004. DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. Em Ana L.C. Bazzan e Sofiane Labidi, editores, *Advances in Artificial Intelligence. XVII Brazilian Symposium on Artificial Intelligence (SBIA'04)*, Lecture Notes in Computer Science, pp. 224–234, Berlin/Heidelberg, 29 de Setembro - 1 de Outubro, 2004. Springer Verlag.

Pardo, Thiago A. S. e Lucia H. M. Rino. 2002. DMSumm: Review and Assessment. Em Nuno Mamede e Elisabete Ranchhod, editores, *Advances in Natural Language Processing: Third International Conference, Proceedings (PorTAL 2002)*, Lecture Notes in Artificial Intelligence, pp. 263–274, Berlin/Heidelberg, 23-26 de Junho, 2002. Springer.

Pekar, Viktor e Richard Evans. 2007. Discovery of language resources on the web: Information extraction from heterogeneous documents. *Literary and Linguistic Computing*, 22(3):329–343.

Peters, Carol, Valentin Jijkoun, Thomas Mandl, Henning Müller, Doug W. Oard, Anselmo Peñas, Vivien Petras, e Diana Santos, editores. 2008. *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF*

2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers, volume 5152 of *Lecture Notes in Computer Science*. Springer, Berlin.

Roberts, Kirk e Andrew Hickl. 2008. Scaling answer type detection to large hierarchies. Em *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. ELRA, 28-30 Maio, 2008.

Rocha, Paulo e Diana Santos. 2007. CLEF: Abrindo a porta à participação internacional em avaliação de RI do português. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, pp. 143–158.

Rocha, Paulo Alexandre e Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Em Maria das Graças Volpe Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, pp. 131–140, São Paulo, 19-22 de Novembro, 2000. ICMC/USP.

Rolo, Carlos Juzarte e António Joaquim Serralheiro. 2008. An approach to natural language equation reading in digital talking books. Em António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, e Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume 5190. Springer Verlag, pp. 268–271.

Santos, Diana. 1995. On grammatical translationese. Em Kimmo Koskenniemi, editor, *Short papers presented at the Tenth Scandinavian Conference on Computational Linguistics*. pp. 59–66, 29-30 de Maio, 1995.

Santos, Diana. 1999a. Porquê processamento computacional do português e não processamento de linguagem natural?, 24 de Março, 1999. `http://www.linguateca.pt/branco/Porque.html`.

Santos, Diana. 1999b. 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/index.html`.

Santos, Diana. 1999c. Towards language-specific applications. *Machine Translation*, 14(2):83–112, Junho, 1999.

Santos, Diana. 2000. O projecto Processamento Computacional do Português: Balanço e perspectivas. Em Maria das Graças Volpe Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*. ICMC/USP, São Paulo, pp. 105–113, 19-22 de Novembro, 2000.

Santos, Diana. 2002a. DISPARA, a system for distributing parallel corpora on the Web. Em Nuno Mamede e Elisabete Ranchhod, editores, *Advances in Natural Language Processing: Third International Conference, Proceedings (PorTAL 2002)*, Lecture Notes in Artificial Intelligence, pp. 209–218, Berlin/Heidelberg. Springer.

Santos, Diana. 2002b. Um centro de recursos para o processamento computacional do português. *DataGramaZero - Revista de Ciência da Informação*, 3(1), Fevereiro, 2002.

Santos, Diana. 2003a. Relatório Linguateca 2000-2003. Relatório técnico, Linguateca, Setembro, 2003. `http://www.linguateca.pt/documentos/RelatorioLinguateca2000-2003Revisto.pdf`.

Santos, Diana. 2003b. Timber! Issues in treebank building and use. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003*, pp. 151–158, Berlin/Heidelberg. Springer.

Santos, Diana. 2004. Aonde vamos em relação a aonde. *the ESPecialist*, 25(1):85–103.

Santos, Diana. 2005. Relatório da Linguateca de 15 de Maio de 2004 a 14 de Maio de 2005. Relatório técnico, Linguateca, 2 de Junho, 2005. `http://www.linguateca.pt/documentos/RelatorioLinguatecaMaio2005.pdf`.

Santos, Diana. 2006a. Desenho, construção e utilização de corpora, 10 de Julho, 2006. `http://www.linguateca.pt/escolaverao2006/Corpora/CorporaEscolaVerao.pdf`.

Santos, Diana. 2006b. Resumo da actividade da Linguateca de 15 de Maio de 2003 a 15 de Dezembro de 2006. Relatório técnico, Linguateca, Dezembro, 2006. Com a colaboração (por ordem alfabética) de Alberto Simões, Ana Frankenberg-Garcia, Belinda Maia, Luís Costa, Luís Miguel Cabral, Luís Sarmento, Marcirio Chaves, Mário J. Silva, Nuno Cardoso, Paulo Gomes e Rui Vilela, `http://www.linguateca.pt/documentos/RelatorioLinguateca2003-2006.pdf`.

Santos, Diana. 2007a. Avaliação conjunta. Em Diana Santos, editor, *Avaliação conjunta: um*

novo paradigma no processamento computacional da língua portuguesa. IST Press, Lisboa, Portugal, pp. 1–12, 20 de Março, 2007.

Santos, Diana, editor. 2007b. *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal.

Santos, Diana. 2007c. Computational linguistics beyond the processing of english. `http://www.linguateca.pt/Diana/download/FirstWords2007.pdf`.

Santos, Diana. 2008a. Curso avançado de estudos contrastivos usando o COMPARA como ferramenta, 3-5 de Novembro, 2008. Módulo na EBraLC, Segunda Escola Brasileira de Linguística Computacional, `http://www.linguateca.pt/documentos/cursoCOMPARASantosEBRALC2008.pdf`.

Santos, Diana. 2008b. Linguateca 10 anos: festejo ou luto?, 11 de Setembro, 2008. `http://www.linguateca.pt/Linguateca10anos/Apresentacoes/AprSantosL10.pdf`.

Santos, Diana. 2008c. Perfect mismatches: Result in English and Portuguese. Em Margaret Rogers e Gunilla Anderman, editores, *Incorporating Corpora: The Linguist and the Translator*. Multilingual matters, Clevedon, pp. 217–242.

Santos, Diana e Anabela Barreiro. 2004. On the problems of creating a consensual golden standard of inflected forms in. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, pp. 483–486, 26-28 de Maio, 2004.

Santos, Diana e Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. Em Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis, e Gregory Stainhauer, editores, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pp. 205–210, 31 de Maio - 2 de Junho, 2000.

Santos, Diana, Luís Miguel Cabral, e Luís Costa. 2006. Linguateca: seven years working for the computational processing of Portuguese, 23 de Novembro, 2006. `http://www.linguateca.pt/Diana/download/AprLinguatecaNov2006.pdf`.

Santos, Diana e Nuno Cardoso. 2005. Portuguese at CLEF 2005: Reflections and Challenges. Em Carol Peters, editor, *Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop (CLEF 2005)*, pp. s/pp, Viena, Áustria, 21-23 de Setembro, 2005. Centromedia.

Santos, Diana e Nuno Cardoso, editores. 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca.

Santos, Diana, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, e Yvonne Skalban. 2009. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. Em Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, e Viviane Petras, editores, *Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Springer.

Santos, Diana e Luís Costa. 2005. A Linguateca e o projecto 'Processamento Computacional do português'. *Terminómetro*, 7:63–69. Número especial - A terminologia em Portugal e nos países de língua portuguesa em África.

Santos, Diana e Luís Costa. 2007. QolA: fostering collaboration within QA. Em Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, e Maximilian Stempfhuber, editores, *Evaluation of Multilingual and Multimodal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers*, volume 4730 of *Lecture Notes in Computer Science*, pp. 569–578, Berlin / Heidelberg. Springer.

Santos, Diana, Luís Costa, e Paulo Rocha. 2003. Cooperatively evaluating Portuguese morphology. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003*, pp. 259–266, Berlin/Heidelberg. Springer Verlag.

Santos, Diana e Ana Frankenberg-Garcia. 2007. The corpus, its users and their needs: a user-oriented evaluation of COMPARA. *International Journal of Corpus Linguistics*, 12(3):335–374, Maio, 2007.

Santos, Diana, Cláudia Freitas, Hugo Gonçalo Oliveira, e Paula Carvalho. 2008. Second HAREM: new challenges and old wisdom. Em

António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, e Paulo Quaresma, editores, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume Vol. 5190, pp. 212–215. Springer Verlag.

Santos, Diana e Caroline Gasperin. 2002. Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation. Em Manuel González Rodrigues e Carmen Paz Suarez Araujo, editores, *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. ELRA, Paris, pp. 597–604, 29-31 de Maio, 2002.

Santos, Diana, Belinda Maia, e Luís Sarmento. 2004. Gathering empirical data to evaluate MT from English to Portuguese. Em Lambros Kranias, Nicoletta Calzolari, Gregor Thurmair, Yorick Wilks, Eduard Hovy, Gudrún Magnúsdóttir, Anna Samiotou, e Khalid Choukri, editores, *Proceedings of LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora*. pp. 14–17, 25 de Maio, 2004.

Santos, Diana e Paulo Rocha. 2005. The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. Em Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, e Bernardo Magnini, editores, *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, volume 3491 of *Lecture Notes in Computer Science*. Springer, Berlin/Heidelberg, pp. 821–832.

Santos, Diana e Luís Sarmento. 2003. O projecto AC/DC: acesso a corpora/disponibilização de corpora. Em Amália Mendes e Tiago Freitas, editores, *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002)*, pp. 705–717, Lisboa, 2-4 de Outubro de 2002, 2003. APL.

Santos, Diana, Rosário Silva, e Susana Inácio. 2008. What's in a colour? Studying and contrasting colours with COMPARA. Em *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pp. s/pp. European Language Resources Association (ELRA), 28-30 de Maio, 2008.

Santos, Diana, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela, e Susana Afonso. 2004. Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa. Em Guillermo De Ita Luna, Olac Fuentes Chávez, e Mauricio Osorio Galindo, editores, *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", IX Iberoamerican Conference on Artificial Intelligence (IBERAMIA 2004)*, pp. 147–154, Novembro, 2004.

Sarmento, Luís, Anabela Barreiro, Belinda Maia, e Diana Santos. 2007. Avaliação de Tradução Automática: alguns conceitos e reflexões. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, pp. 181–190.

Sarmento, Luís e Belinda Maia. 2003. Gestor de corpora - Um ambiente Web integrado para Linguística baseada em Corpora. Em José João Almeida, editor, *Corpora Paralelos, Aplicações e Algoritmos Associados (CP3A)*, pp. 25–30, Braga, 3 de Junho, 2003. Universidade do Minho.

Sarmento, Luís, Belinda Maia, e Diana Santos. 2004. The Corpógrafo - a Web-based environment for corpora research. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*. pp. 449–452, 26-28 de Maio, 2004.

Seco, Nuno e Nuno Cardoso. 2006. Detecting user sessions in the tumba! web log. Relatório técnico, Linguateca, Março, 2006. `http://eden.dei.uc.pt/~nseco/tumba.pdf`.

Seco, Nuno, Diana Santos, Rui Vilela, e Nuno Cardoso. 2006. A Complex Evaluation Architecture for HAREM. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira, e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006*, volume LNAI 3960, pp. 260–263, Berlin/Heidelberg. Springer Verlag.

Serralheiro, A., I. Trancoso, D. Caseiro, T. Chambel, L. Carriço, e N. Guimarães. 2003. Towards a repository of digital talking books. Em *EUROSPEECH 2003 - 8th European Conference on Speech Communication and Technology (Interspeech'2003)*. Genebra, Suíça, Setembro, 2003.

Silva, Augusto Soares. 2008a. Integrando a variação social e métodos quantitativos na investigação sobre linguagem e cognição: para uma sociolinguística cognitiva do português europeu e brasileiro. *Revista de Estudos da Linguagem*, 16(1):49–81.

Silva, Mário J. 2008b. Pólo XLDB da Linguateca: 4 anos, 11 de Setembro, 2008. Apresentação no Encontro Linguateca: 10 anos, `http://www.linguateca.pt/Linguateca10anos/Apresentacoes/AprMJSilvaL10.pdf`.

Simões, Alberto. 2008. *Extracção de Recursos de Tradução com base em Dicionários Probabilísticos de Tradução*. Tese de doutoramento, Faculdade de Engenharia da Universidade do Minho, Braga, Março, 2008.

Simões, Alberto e José João Almeida. 2007. Parallel Corpora based Translation Resources Extraction. *Procesamiento del Lenguaje Natural*, 39:265–272, Setembro, 2007.

Vallin, Alessandro, Bernardo Magnini, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Simov, e Richard Sutcliffe. 2005. Overview of the CLEF 2004 Multilingual Question answering track. Em Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, e Bernardo Magnini, editores, *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, volume 3491 of *Lecture Notes in Computer Science*. Springer, Berlin/Heidelberg, pp. 371–391.

Vilela, Rui, Alberto Manuel Simões, Eckhard Bick, e José João Almeida. 2005. Representação em XML da Floresta Sintáctica. Em José Carlos Ramalho, Alberto Simões, e João Correia Lopes, editores, *3ª Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas (XATA 2005)*, pp. 351–361. Departamento de Informática, Universidade do Minho.

Wing, Benjamin e Jason Baldrige. 2006. Adaptation of Data and Models for Probabilistic Parsing of Portuguese. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira, e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR'2006)*, volume LNAI 3960, pp. 140–149, Berlin/Heidelberg. Springer.

Xavier, Maria Francisca, Maria de Lurdes Crispim, Graça Vicente, A. Castro, Alexandra Fiéis, Maria Cristina Silva, e M. Lobo. 1998. Utilizações informáticas de corpora textuais medievais. Em Palmira Marrafa e Maria Antónia Mota, editores, *Linguística Computacional: Investigação Fundamental e Aplicações. Actas do I Workshop sobre Linguística Computacional da Associação Portuguesa de Linguística*. Colibri, Lisboa, pp. 347–358.