Disponibilização de *corpora* de texto através da WWW

Diana Santos
Processamento computacional do português
SINTEF Telecom and Informatics

Neste artigo pretendo, após a minha experiência com o Oslo Corpus of Bosnian Texts (Santos, 1998),

1. Discutir as vantagens de tornar acessíveis *corpora* de texto[1] através da rede
2. Analisar possíveis entraves à sua disponibilização em Portugal
3. Sugerir uma colaboração inter-grupos no sentido de disponibilizar *corpora* portugueses

Assumo como dado adquirido que as vantagens de poder recorrer a *corpora* de texto são indiscutíveis quando se estuda a língua, se desenvolvem ferramentas para o seu tratamento computacional, ou se criam recursos para estudantes, professores, e a comunidade dos falantes em geral (como é o caso de dicionários, gramáticas, prontuários, etc.). Este artigo versará portanto apenas as vantagens de os *corpora* estarem disponíveis a todos quantos se dedicam à engenharia de linguagem.

Corpora de texto

Um *corpus* é geralmente considerado como uma colecção de textos cuja compilação não foi arbitrária. Por vezes distingue-se entre banco de textos e *corpus* propriamente dito. Exactamente quais os atributos que um *corpus* deve ter não é, contudo, objecto de consenso; além disso, mesmo quando os investigadores concordam sobre alguns requisitos a que um *corpus* ideal deva obedecer, discordam em geral em relação à forma prática de o conseguir. (Por exemplo, o critério da representatividade.)

Na área de "linguística com corpora" (‘corpus linguistics’) – área em que os investigadores consideram os *corpora* como centrais para a investigação linguística, e que, portanto, em paralelo com os problemas linguísticos que constituem o objecto da sua pesquisa, se debruçam sobre os meios que lhes permitam prossegui-la – existem, *grosso modo*, dois tipos de grupos/pessoas:

- os compiladores, que se preocupam sobretudo com questões de desenho dos *corpora*, revisão e anotação dos mesmos, comparação entre vários *corpora*
- os utilizadores, que se preocupam sobretudo com a informação extraível de quaisquer *corpora* em que possam tocar[2]

[1] Visto que a situação é consideravelmente diferente na área da fala, remeto o leitor para Martins et al. (1998).

[2] Um terceiro grupo, o dos desenhadores de ferramentas de *corpora*, começa também a tornar-se notado, como as recentes discussões na lista electrónica CORPORA mostram (veja-se o arquivo das mensagens a esta lista, em http://www.hit.uib.no/corpora/). Exemplos conhecidos são Oliver Christ (Christ, 1994), Lou Burnard (Aston & Burnard, 1996) e, em Portugal, José Carlos Medeiros (Medeiros, 1992).

Ainda que tal fenómeno seja evidente na prática, os investigadores não se vêem em geral apenas com um destes perfis. Como tal, todos os compiladores «estão a compilar o corpus para depois o poderem usar, naturalmente», e os utilizadores muitas vezes têm de compilar os seus próprios *corpora* (por falta de recursos externos) ou usar aquilo que lhes vem à mão (e que, em geral, não obedece aos requisitos mínimos que os compiladores postulam para ser chamado um *corpus*).

Existe pois, no meu entender, um grande subaproveitamento dos recursos (físicos e humanos) envolvidos na linguística com *corpora*: em geral, os *corpora* mais bem desenhados são os menos utilizados (porque os recursos humanos a eles afectos estão sobretudo envolvidos na compilação, preparação e revisão do material – que é bem morosa e delicada); por outro lado, a maior parte dos estudos efectuados utilizando *corpora* recorrem a material pouco tratado.

Além disso, e a uma dimensão mais vasta, a maior parte da comunidade linguística, não tendo acesso aos *corpora*, não desenvolve aptidões para explorar e raciocinar com base neste tipo de recursos, não sentindo pois sequer a necessidade de se tornar um seu utilizador.

A maneira de contrariar esta tendência infeliz, visível a nível internacional mas ultrapronunciada no nosso país, é disponibilizando os *corpora* à medida que forem sendo criados (e não ficar à espera da última moda); e premiando, em vez de prejudicar, os compiladores, por cada novo utilizador que consigam arranjar.

Uma das maiores vantagens da criação de *corpora* que estejam disponíveis publicamente é a possibilidade de medir (avaliar) uma dada ferramenta, ou teoria, em relação a um padrão comum. Tal vantagem, listada sempre em primeiro lugar na literatura dos *corpora* falados – veja-se, para o português, Martins et al. (1998) —, é raras vezes sublinhada no contexto dos *corpora* escritos, de línguas diferentes do inglês. Contudo, o Brown corpus (Francis & Kucera, 1979) e, em menor grau, o LOB corpus (Johansson et al., 1978), foram durante anos a medida para avaliar anotadores (‘taggers’) do inglês.

Esta é, de facto, uma das grandes diferenças entre a linguística com *corpora* e a linguística de introspecção (em que as intuições de gramaticalidade e naturalidade são próprias de cada linguista envolvido, e geralmente em variação evidente com as dos proponentes da teoria alternativa).[3] Ao produzir uma base comum de estudo, permite-se tornar mais objectivo[4] o progresso na área das ciências humanas e sobretudo na área de criação de ferramentas que processem a língua, assim como na criação de bancos de dados sobre a língua.

---

[3] Convém referir que nenhum proponente da linguística com *corpora* pretende excluir a competência introspectiva do analista, e que portanto um *corpus* não substitui a instrospecção.

[4] Note-se que um *corpus* contém fenómenos pelo menos tão marginais como os criados pela instrospecção dos linguistas teóricos – a diferença é que tais exemplos não são *a priori* convenientes para nenhuma teoria...

A disponibilização de *corpora* não é, contudo – ou, pelo menos, a curto prazo – uma panaceia para pôr os investigadores da língua portuguesa nas primeiras fileiras da linguística com *corpora* e o português no pelotão da frente das línguas bem tratadas computacionalmente. Sabe-se, de outras experiências, que não é só dizer «está aqui» e todos sem excepção passam a usar o *corpus* na sua investigação, mudando radicalmente os seus hábitos de trabalho e os seus métodos de raciocínio.[5]

De facto, disponibilizar os *corpora* não chega: muitos utentes, senão a maior parte, após as primeiras procuras não saberão o que fazer. A linguística com *corpora* não é só (nem sobretudo) a arte de criar *corpora*, é sim a arte de os utilizar para o avanço da linguística em geral e do processamento de uma dada língua em particular.

Muitos, contudo – pelo menos os grupos envolvidos em processamento de linguagem natural no nosso país –, já teriam utilização imediata para tal disponibilização / partilha de recursos. Mas com muito raras excepções (o projecto Natura (Almeida et al., s/d) e, a uma dimensão muito mais reduzida, o CorpusINESC (Santos, 1992)), os *corpora* não estão disponíveis fora das instituições em que foram compilados.[6]

É esta situação que sugiro contrariar, e inverter, criando, a médio prazo, uma rede de recursos do português que permita que a comunidade beneficie do património da nossa língua e tenha mais recursos para fazer investigação, desenvolvimento e teste de produtos.

Mas para mudar uma situação que se reputa má, é preciso começar por compreender as suas causas. Ou seja, porque é que os recursos (em particular os *corpora*) não se encontram disponíveis? Começo, pois, por fazer aqui uma breve descrição dos entraves à disponibilização:

1. Questões legais e de lucro

Muitos dos compiladores de *corpora* não têm experiência em questões legais. Por vezes, são demasiado cautelosos nos pedidos que fazem aos donos dos textos: note-se que se fizerem um contrato em que apenas eles podem usar os textos, mais tarde não é possível disponibilizá-los.[7] Outras vezes, no extremo oposto, alguns utilizadores de *corpora* usam textos / material para o qual não se

---

[5] De facto, a minha experiência pessoal leva-me a crer que não basta dar ferramentas, é preciso educar as pessoas no seu uso.

[6] Estou-me a referir aqui apenas aos *corpora* criados em Portugal. Existem outros *corpora* acessíveis que incluem o português, mas que não foram compilados com participação portuguesa, como por exemplo os corpora ECI-MCI, MLCC, LDC95T11 (AFP). Veja-se http://www.portugues.mct.pt/recursos.html.

[7] Quero aqui citar o apelo feito por Mark Libermann, presidente do Linguistic Data Consortium (LDC, veja-se http://www.ldc.upenn.edu/) após conversa sobre *corpora* portugueses: "I would also hope that you would work to obtain general distribution rights, that is, the right to distribute to all interested researchers around the world, rather than just to a fixed group of Portuguese institutions, or some other restricted group such as academics (as in the case of the LOB corpus) or Europeans (as in the case of the BNC)."

preocuparam em obter direitos. Por isso, mais tarde também não podem tornar o material acessível sem incorrer em problemas legais.

Por outro lado, muitas pessoas pensam que um *corpus* é, à partida, algo que pode valer muito, e que é preciso por isso ter muito cuidado, evitando que outros retirem lucro do seu trabalho. Penso que na maioria das vezes essas pessoas não têm razão. Ou que, pelo menos, a questão é muito mais complicada. De facto, um *corpus* por si só não tem valor. Mesmo fazer um dicionário ou uma lista de frequências baseado nele não é trivial (e, além disso, se fosse só baseado nele seria incompleto), e esses seriam os produtos mais directamente "deriváveis" de um *corpus*. De facto, as vantagens mais imediatas seriam o teste, afinação, e melhoria de um produto, ou dicionário, que já existisse na prática. Seria, pois, difícil de quantificar a contribuição do uso do *corpus*. Contudo, um argumento que me parece muito importante é que tal uso também poderia trazer valor acrescentado ao próprio *corpus*. (Por exemplo, a detecção de incorrecções, informação sobre neologismos e, claro, a adição de informação sobre o próprio *corpus*.) Na prática, na maior parte das vezes seria mais fácil e vantajoso tentar partilhar os lucros com os utilizadores e não evitá-los.[8] De facto, são em geral as comunidades científica e universitária que se dedicam à criação de corpora, o que faz com que o não façam numa óptica empresarial, e que portanto seja de esperar que distribuam o resultado dos seus trabalhos, em vez de o considerarem um produto a nível de mercado.

No entanto, são muitas vezes também os donos dos textos (editoras, jornais, empresas públicas) que fazem questão em restringir o acesso, quer por sobrevalorização do seu produto (textos)[9], quer por ignorância em relação ao que está, de facto, em causa. Convém dizer que este não é necessariamente um mal nacional: em vários outros países questões análogas aparecem ou apareceram; mas surgiram também várias iniciativas para evitar uma preocupação excessiva com os direitos de autor. A comunidade portuguesa devia, na minha opinião, envolver-se numa clarificação da situação antes que seja demasiado tarde.

2. Questões técnicas

Em muitos casos, os compiladores de corpora não têm capacidade técnica, financeira e humana para disponibilizar os recursos que só eles sabem utilizar. Disponibilizar, ou mesmo apenas distribuir,

---

[8] Esta observação não se aplicaria se o *corpus* fosse criado por uma editora que já possuísse os produtos que queria melhorar. Contudo, é importante notar que no nosso país os *corpora* existentes foram todos criados por organizações não lucrativas... E, mesmo nos EUA, a criação do American National Corpus, sugerida recentemente em Granada, apela às agências financiadoras, assim como afirma que nenhuma instituição americana, sozinha, poderá levar a cabo tal mega-tarefa (Fillmore et al., 1998).

[9] Note-se que se está a falar, neste contexto, de extractos e não de uma obra completa, e que não se está sugerir que os donos dos direitos legais não sejam eventualmente compensados de uma forma ou de outra.

significa tempo, esforço, documentação, serviço aos utilizadores.[10] Este tipo de actividades é raramente planeado, previsto, financiado, mesmo escolhido pelos investigadores. Contudo, é algo que é essencial se não se quer deixar o trabalho morrer sem ter sido (bem) utilizado.

Por outro lado, é preciso salientar que, mesmo com todos os recursos e excelentes intenções de ajudar o próximo, para desenhar e implementar um serviço de WWW é preciso competência técnica específica que em geral não existe nos grupos das diversas áreas da linguística com *corpora*, e em particular naqueles que se dedicam à compilação de *corpora*.

É, por isso, crítico, não só que haja colaboração entre grupos com perfis diferentes, mas que as pessoas que trabalhem em processamento de linguagem natural desenvolvam capacidades quer de análise linguística quer de processamento computacional.

3. Questões de mentalidades

Até agora, tem sido visto como uma vantagem possuir um bem / recurso que os outros grupos não possuam, e tal tem sido apresentado como um argumento para renovação (ou obtenção) de financiamento. Se isto é possivelmente o caso num contexto empresarial, não é, de facto, na minha opinião, um argumento válido no caso da investigação, cujo financiamento não devia proteger grupos nem favorecer recursos proprietários, mas sim recursos partilháveis e utilizáveis por todos. É também discutível que se deva negar o acesso a esses recursos no caso de empresas nacionais.

É interessante notar que, no contexto da avaliação em engenharia da linguagem, tem sido muito discutido o posicionamento e atitude dos diversos grupos de investigação, tendo sido sugerido que os programas de avaliação comuns nos Estados Unidos deram origem a um clima de "colaboração competitiva" benéfico para a comunidade científica. E, de facto, é também neste país que a proposta de criar um corpus nacional é feita, de raiz, num modo colaborativo, sugerindo que toda a comunidade científica se deva unir para o obter, dividindo tarefas e recursos (humanos, técnicos e financeiros).[11]

---

[10] Este é uma das questões em que insisto em Santos (1998), e que era também um *leitmotiv* na conferência LREC (Language Resources and Evaluation) em Granada, Maio 1998. A manutenção (ou desenvolvimento contínuo) dos recursos foi considerada como essencial e a sua omissão dos programas de financiamento um dos piores males para o processamento de linguagem natural, como exposto em Macleod (1998).

[11] Como aliás não seria de admirar, não pretendem restringir o seu uso apenas a investigação, nem pretendem afastar empresas: pelo contrário, sugerem a criação de um consórcio de editoras como parceiro integrante, à imagem do British National Corpus (BNC, veja-se http:// http://info.ox.ac.uk/bnc/).

Corpora acessíveis pela WWW

Longe de querer fazer aqui uma apresentação da WWW, pretendo apenas indicar quais as vantagens de a usar para disponibilizar *corpora*. Algumas vantagens advêm directamente das propriedades da própria rede:

1. A sua divulgação. É indesmentível o uso cada mais generalizado da Internet no nosso país, sobretudo a nível das camadas mais jovens. A Web é cada vez mais acessível a um número maior de pessoas, e "acessível" não quer dizer apenas fisicamente mas sobretudo intelectualmente acessível. (As pessoas aprendem a manipulá-la nas escolas, nos empregos, em casa.)
2. A segunda vantagem é que torna possível que utilizadores individuais acedam a grandes acervos de informação sem terem necessariamente um computador poderoso e com muita memória. Ou seja, do ponto de vista de um utilizador privado, pode aceder-se aos computadores de outras pessoas e instituições; do ponto de vista das organizações, pode-se distribuir a informação sem ter de a centralizar.

Mais especificamente sobre a disponibilização de *corpora* pela rede, temos as seguintes vantagens do ponto de vista da instituição que possui ou compilou o *corpus*:

3. É possível restringir o acesso de forma a evitar (se for esse o objectivo) que o utilizador receba o *corpus* inteiro no seu computador.
4. É possível saber o que os utilizadores fazem com os *corpora*, monitorizando o acesso.
5. É possível corrigir/melhorar o *corpus* sem que vários utilizadores usem versões diferentes.
6. É possível contabilizar o acesso, e o trabalho feito sobre o *corpus*, e daí eventuais lucros ou, mais vagamente, a eventual resposta da comunidade científica e dos parceiros comerciais.

E, do ponto de vista do utilizador, temos os seguintes benefícios:

7. o acesso, do seu local de trabalho ou casa, a um recurso que se encontra fisicamente longe;
8. a minimização dos conhecimentos técnicos necessários para aceder ao *corpus*: não é preciso instalar programas; mudar de sistema operativo; aprender uma sintaxe complicada para conseguir chegar a "pôr as mãos" no *corpus*;[12]
9. a minimização dos recursos tecnológicos necessários (espaço de memória, requisitos de sistema operativo) – basta ter um "folheador" ('browser') da rede;
10. a existência de um apoio remoto e de uma comunidade de outros utilizadores com quem pode trocar experiências, resolver problemas, e colaborar.

---

[12] Está claro que será sempre preciso aprender como procurar aquilo em que se está interessado; e de facto, saber como refinar as procuras num corpus, e como avaliar os resultados do ponto de vista da confirmação ou infirmação de uma dada hipótese não é simples, e constitui um dos aspectos mais interessantes da linguística com *corpora*. Este tipo de questões não se prende, contudo, com a dificuldade meramente sintáctica de fazer o sistema obedecer aos nossos comandos, que é ao que me estou a referir aqui.

A experiência com o Oslo Corpus of Bosnian Texts provou que estas vantagens não são simples hipóteses académicas. Consequentemente, e com menos problemas técnicos (a questão dos caracteres portugueses é mais simples do que a dos bósnios, visto que está incorporada no padrão ISO-Latin-1, ao invés de ISO-Latin-2), seria possível construir vários destes serviços para os *corpora* existentes em Portugal ou nos grupos que processam o português.[13] Esta rede de recursos permitiria não só o acesso da comunidade portuguesa ao trabalho e dados dos outros grupos, contornando eventuais problemas mais complicados de direitos de autor (visto que só se daria acesso para consulta, e não para apropriação ou cópia do *corpus* inteiro). Além disso, potenciaria eventuais colaborações, comparações e avanços surgidos da cooperação, como seria o caso da produção de *corpora* anotados e de interfaces mais poderosas de acesso aos mesmos, que poderiam ser partilhados pela comunidade.

Desde já aqui me ofereço para partilhar a experiência e eventuais programas de disponibilização com os grupos que possuem *corpora* de texto português, de forma a que, a curto prazo, seja possível disponibilizar o acesso aos *corpora*, em rede, de preferência com mais informação do que o simples conteúdo lexical.

Com efeito, é perfeitamente exequível neste momento, do ponto de vista técnico, a disponibilização de *corpora* em português com informação sobre categoria sintáctica e texto de origem, com a tecnologia presente e uma colaboração mínima entre vários grupos envolvidos no processamento do português. Seria também desejável que a curto prazo se fomentasse a criação de *corpora* analisados, para permitir uma avaliação independente e fiável dos analisadores para o português que se pretendam implementar ou que já existam.

Finalmente, como última proposta neste artigo "político" (por falta de espaço, remeto o leitor para os pormenores técnicos da disponibilização de *corpora* na WWW dados em Santos (1998)), sugiro que, em paralelo com a criação da rede de recursos descrita antes, se lance um programa de avaliação / competição amigável entre vários grupos, à semelhança dos Evaluation Contests (que sugiro chamar Avaliações Conjuntas) patrocinados pela National Science Foundation (NSF) dos Estados Unidos, em torno de objectivos concretos e objectivamente observáveis, no processamento do português.[14] Tais objectivos seriam pré-definidos pela comunidade científica, que aceitaria a classificação resultante da competição como (um) motivo de financiamento. O simples facto de todos os participantes se terem de envolver na definição de um método comum de avaliação iria levar a um maior consenso sobre quais as questões difíceis no tratamento computacional da nossa língua, e quais

[13] Veja-se a recente catalogação dos recursos, em termos de *corpora* existentes para o português mencionados na WWW, em http://www.portugues.mct.pt/recursos.html#corpora.

[14] Esta ideia não é obviamente minha, nem sequer originalmente para o português. Foi-me sugerida (pelo menos) pela Isabel Trancoso e pelo recente relatório estratégico de Hovy et al. (1998).

as pontes possíveis entre teorias e opiniões diferentes – tudo na óptica, é claro, de uma dada aplicação ou objectivo prático.

O facto de haver vários *corpora* e recursos textuais criados por grupos diferentes seria, neste contexto, uma garantia de que nenhum grupo teria, à partida, vantagens por afinar / ter afinado um sistema no seu *corpus* (que, a ser o único, conteria também o *corpus* de teste sobre o qual a avaliação iria decorrer).

Agradecimento

Agradeço à Signe Oksefjell várias sugestões para a melhoria do presente artigo.

Referências

Almeida, José João Dias de, José Carlos Ramalho, Ulisses Pinto & Ricardo Reis. "Projecto Natura", http://www.di.uminho.pt/~jj/pln/pln.html, sem data.

Aston, Guy & Lou Burnard. *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, 1996.

Christ, Oliver. "A modular and flexible architecture for an integrated corpus query system", *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research* (Budapest, July 7-10, 1994), pp.23-32.

Fillmore, Charles, Nancy Ide, Dan Jurafsky & Catherine Macleod. "An American National Corpus: A Proposal", in Rubio et al. (1998), Vol. 2, pp.965-9.

Francis, W.N. & Kucera, H. *Manual of information to accompany a Standard Sample of Present-day Edited American English, for use with digital computers*, 3rd edition, 1979 [1964].

Hovy, Eduard, Nancy Ide, Robert Frederking, Joseph Mariani & Antonio Zampolli. "Multilingual Information Management: Current Levels and Future Abilities", http://www.cs.cmu.edu/ ~ref/mlim/, Julho 1998.

Johansson, S., G. Leech & H. Goodluck. *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*, Oslo, 1978.

Martins, Ciro, Isabel Mascarenhas, Hugo Meinedo, João Neto, Luís Oliveira, Carlos Ribeiro, Isabel Trancoso & Céu Viana. "Spoken Language Corpora for Speech Recognition and Synthesis in European Portuguese", *Proc. RECPAD'98 - 10th Portuguese Conference on Pattern Recognition* (Lisboa, Março 1998), http://www.speech.inesc.pt/bib/Trancoso98a/ poster.html

Macleod, Catherine. "A Plea for Consideration of Maintenance of Language Resources", in Rubio et al. (1998), Vol. 1, pp.35-40.

Medeiros, José Carlos. "Ferramentas de processamento de corpora usando o PALAVROSO", in Santos (1992).

Rubio, Antonio, Natividad Gallardo, Rosa Castro & Antonio Tejada (eds.). *Proceedings of The First International Conference on Language Resources and Evaluation* (Granada, 28-30 May 1998).

Santos, Diana (ed.). "Processamento de corpora de texto no INESC", Relatório INESC nº. RT/65-92, Dezembro, 1992.

Santos, Diana. "Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts", in Rubio et al. (1998), Vol. 1, pp.475-81.