Comparação de corpora em português: algumas experiências

Diana Santos

Neste artigo dou os primeiros passos na comparação de corpora em português, utilizando para tal a maioria dos recursos textuais disponíveis publicamente, e apresentando alguns resultados preliminares.

1. Motivação

No momento presente assiste-se a uma quase unanimidade em processamento de linguagem natural (ou linguística computacional). Todos concordam que é muito importante ter acesso a recursos textuais: corpora de texto, ainda que possivelmente alguns não os saibam sequer usar. Fala-se muito na criação de recursos de referência que sigam formatos e regras que facilitam a sua troca, fala-se também muito em questões de direitos e de disponibilização, e no valor acrescentado que a compilação dos mesmos representa em relação ao uso simples de textos. Fala-se, infelizmente, muito pouco no próprio uso desses corpora para tarefas concretas e ainda menos na sua avaliação e comparação. Esta situação é particularmente flagrante no que se refere a corpora de texto em língua portuguesa.

A questão que se coloca (e que ainda não foi respondida) é se, de facto, vale a pena investir em corpora genéricos, para todos os usos, que sejam vistos como repositório de informação a ser partilhada por quantos se dedicam ao processamento de uma língua; ou se não será mais eficiente, económico e útil produzir (ou recorrer a) corpora diferentes para cada uma das aplicações que se tem em mente.

Um primeiro passo na investigação dessa questão é tentar comparar todos os corpora a que é possível ter acesso actualmente (e eventualmente criar um corpus-união). Relatando algumas experiências de comparação de corpora de português, o artigo produz uma descrição relativamente pormenorizada de cada corpus que já existe no domínio público, experimentando metodologias que poderão mais tarde ser aplicadas a outros corpora (ou utilizadas no seu desenho).

Um problema fundamental para a comunidade que se dedica ao processamento computacional da língua portuguesa é, obviamente, a questão das diferenças entre as normas portuguesa e brasileira, cuja sistematização parcial através do estudo de corpora já foi proposta, por exemplo, por Wittmann, Pêgo, & Santos (1995).

Para poder avaliar as diferenças, é preciso saber qual a variabilidade expectável dentro de cada variante e mesmo dentro de um dado corpus (Kilgarriff, 1997b). É este problema que o presente artigo pretende ajudar a resolver, ao apresentar alguns estudos preliminares nesse sentido.

2. Método

Para poder fazer algumas experiências, por modestas que sejam, é preciso converter todo o material para um mesmo formato. Saliente-se desde já que não se fala aqui em representar a mesma informação, porque corpora diferentes contêm informação diferente. Este problema, inescapável, foi resolvido de forma diferente – muito mais ambiciosa – por Peters et al. (1998) no que se refere a recursos lexicais.

No nosso caso apenas codificámos todos os corpora no mesmo sistema de processamento de

corpora, tentando sempre que possível usar os mesmos atributos estruturais. O formato usado foi o do IMS Corpus Workbench (veja-se Christ et al. (1999) para uma descrição detalhada das suas potencialidades e Santos & Ranchhod (1999) para uma discussão dos méritos e propriedades deste ambiente para processar o português). Foram depois desenvolvidos alguns programas específicos, enriquecendo a série de funcionalidades associadas ao IMS-CWB.

3. Apresentação dos corpora

Segue-se a lista dos corpora identificados e obtidos a partir do catálogo descrito em Oksefjell & Santos (1998) e acessível de http://www.portugues.mct.pt. A ordem por que são apresentados reflecte apenas a ordem por que foram processados.

- **Natura/Público** Corpus jornalístico Natura-PUBLICO, http://www.di.uminho.pt/~jj/pln/pln.html
- **ENPCport** Parte portuguesa (traduções para português de originais em inglês) do English-Norwegian Parallel Corpus (ENPC), Johansson, Ebeling, & Oksefjell (1999), Oksefjell (1998; 1999)
- **Natura/Minho** Corpus jornalístico Natura-Diário do Minho, http://www.di.uminho.pt/~jj/pln/pln.html
- **ECI-BR** e **ECI-EE** Respectivamente a parte do corpus Borba-Ramsey e a da apresentação do programa Esprit do European Corpus Initiative, the Multilingual Corpus 1 (ECI/MCI), McKelvie & Thompson (1994), Thompson et al. (1994)
- **MLCC-DEB** MLCC - Multilingual Corpora for co-operation, Sub-corpus: Official Journal of the European Commission, Annex: Debates of the European Parliament 1992-1994, Armstrong et al. (1997; 1998)
- **NILC/São Carlos** Corpus NILC/São Carlos (parte corrigida), Nunes et al. (1996a; 1996b)

Note-se que os formatos em que os corpora se apresentavam variavam desde texto simples (fragmentos: Corpus Natura/Público; ou textos completos: Corpus NILC/São Carlos) até incluindo codificação variada do tipo SGML (fragmentos: ENPCport; ou textos completos: MLCC-DEB).

Na Tabela 1 apresento uma primeira[1] descrição do material, remetendo o leitor para http://www.portugues.mct.pt/corpora/contabilizacao.html para contabilizações mais detalhadas.

Corpus	Nº de frases	Nº de parágrafos	Nº de palavras (formas)	Nº de palavras (tipos)	Nº de elementos
Natura/Público	224 855	79 458	6 226 724	172 612	7 223 696
ENPCport	15 531	5 462	231 842	26 586	271 133
Natura/Minho	80 396	71 119	2 112 665	78 799	2 610 551
ECI-EBR	43 553	12 118	713 038	60 153	878 311
ECI-EE	769	391	26 310	4 147	30 019
MLCC-DEB	319 482	108 935	8 867 493	111 142	10 180 570
NILC/São Carlos	2 038 046	793 481	31 563 621	426 487	39 628 743

[1] A manipulação e consequente depuração de um corpus é um processo iterativo, por isso todos os valores aqui apresentados estarão sujeitos a refinamentos à medida que se aperfeiçoem os programas e a própria codificação dos corpora.

Tabela 1: Contabilização básica dos corpora disponíveis

O processamento inicial de cada um destes corpora tentou preservar (ou recuperar) uma divisão em parágrafos e identificar a divisão em frases (apenas já existente no caso do ENPCport; efectuada automaticamente nos outros casos). Por outro lado, toda a informação sobre formatação (por oposição a estrutura textual) foi retirada. Outras divisões, possivelmente correspondentes a uma sintaxe diferenciada (por exemplo títulos, subtítulos, legendas de figuras, notas de rodapé, assinaturas de artigos de jornal ou de cartas, etc.), foram mantidas nos casos em que tal recuperação se pudesse fazer automaticamente. A divisão em unidades narrativas também foi mantida quando presente no corpus (ou seja, a separação entre artigos diferentes, textos diferentes, capítulos diferentes, etc.).

3.1. Algumas questões de codificação

Não é possível criar corpora informatizados sem efectuar decisões subjectivas quanto à sua escolha e armazenamento. Isso também acontece quando se tem que converter, para um mesmo sistema, corpora criados por grupos diferentes com escolhas diferentes.

O tratamento das notas de rodapé em três corpora diferentes, todos usando SGML (a saber, ENPCport, MLCC-DEB e ECI-EE), ilustra claramente que seguir o mesmo padrão de codificação de um corpus não significa necessariamente as mesmas opções. Aliás, convém notar que as próprias notas de rodapé (traduzidas por uma formatação diferenciada) podem possuir objectivos diferentes em textos de estilos diferentes ou mesmo advir simplesmente de preferências individuais ou editoriais distintas.

Daí que também o seu tratamento no âmbito do presente trabalho tenha sido diferente: No caso do ENPCport, a nossa opção foi a de simplesmente eliminar as (33) notas (em geral notas do tradutor) do corpus final – visto que o seu objectivo principal, o de esclarecer ou traduzir palavras ou expressões do texto, não nos pareceu que merecesse o trabalho adicional de pré-processamento necessário para criar o corpus. Nos outros casos, as notas foram identificadas como pertencendo a uma secção estrutural do tipo <nota>, arrumando-a, em termos sequenciais, após o parágrafo em que a chamada à nota é feita. No ECI-EE foi preciso fazer a identicação das notas manualmente, enquanto que no MLCC-DEB apenas foi preciso modificar a sua posição no ficheiro.

4. Comparação

4.1. Propriedades ortográficas

A forma mais trivial de comparar os vários corpora é olhar para os seus pormenores ortográficos, ou seja, contar as formas variadas em que os seus caracteres se apresentam. Para permitir uma comparação mais fácil, apenas os valores relativos (em percentagem (%) ou permilagem (%%)) são apresentados na Tabela 2.[2] O número superior refere-se a tokens, o inferior a tipos, ou seja, no corpus Natura/Público 77,62% das palavras eram totalmente constituídas por letras minúsculas, correspondendo a 51,95% de todas as palavras diferentes encontradas no corpus.

[2] Os valores são calculados através da divisão de cada contagem pelo número total de "palavras" (formas), incluindo estas números, abreviaturas e siglas, mas não incluindo pontuação nem codificação estrutural.

Corpus	Minúsculas (%)	Inicial Maiúscula (%)	Maiúsculas (%)	Minúsculas com hífen (%)	Inicial masc. com hífen (%%)	Maiúsculas com hífen (%%)	Números (%)	Com números (%%)	Pontuação/frase	Palavras/frase
Natura/ Público	77,62 51,95	10,80 26,72	0,70 2,99	0,83 7,50	2,22 35,47	0,16 2,10	1,01 1,07	0,90 8,31	1,73	27,7
ENPCport	79,19 74,51	9,92 14,12	0,10 0,47	1,32 7,66	3,76 22,04	0,02 0,19	0,13 0,51	0,09 0,34	1,17	14,9
Natura/ Minho	71.69 53,08	14,48 23,95	1,06 3,96	0,84 7,12	4,56 41,74	0,16 1,94	1,54 2,12	3,95 7,37	1,95	26,2
ECI-EBR	78,88 69,64	10,15 20,12	0,03 0,08	0,87 6,65	2,45 19,85	~0 0,03	0,24 0,71	0,08 0,68	1,24	16,4
ECI-EE	81,98 80,15	5,66 8,42	1,65 2,56	1,00 3,81	1,10 5,79	0,15 0,96	0,77 1,71	0,95 1,45	1,54	34,2
MLCC-DEB	80,07 54,60	9,23 15,04	0,50 1,55	0,79 11,79	3,64 29,43	0,21 4,67	0,69 1,38	0,74 6,96	1,23	27,8
NILC/São Carlos	71,26 34,89	14,64 29,19	1,79 5,54	0,68 9,03	1,97 45,76	3,56 5,93	1,30 0,91	1,71 13,12		15,5

Tabela 2: Distribuição básica de características gráficas

É interessante notar que os textos literários são os que possuem acentuadamente menos palavras por frase. Tal pode dever-se ao maior comprimento, em formas, das entidades mencionadas em texto jornalístico ou político (veja-se a secção sobre nomes próprios, abaixo) e também à utilização de discurso directo.

Outra constatação interessante é o elevado número de formas com inicial maiúscula em todos os corpora, sendo agora o discurso político aquele que menos destas palavras apresenta. Por outro lado, o número de clíticos em posição enclítica não parece favorecer nenhum dos corpora, o que concorda com a intuição de que esta característica corresponde a uma propriedade nuclear da língua portuguesa (e não especificamente de um género literário). Seria, evidentemente, interessante olhar para o tipo de clíticos, o número de palavras compostas e a quantidade de identificadores incorporando um hífen, para indagar se a este nível as diferenças já seriam significativas.

4.2. Palavras mais frequentes

Outra comparação relativamente fácil de fazer é entre as cem palavras mais frequentes destes corpora. A Tabela 3 apresenta a sublinhado as palavras únicas no grupo das cem primeiras e a itálico as palavras que pertencem ao grupo das cem primeiras em pelo menos todos os corpora menos um (precisamente cinquenta e uma, apresentadas na última linha por ordem alfabética). Para cada corpus as palavras estão ordenadas por frequência decrescente.

Corpus	100 palavras mais frequentes, por ordem
Natura/Público	*de a o que e do da em um para os uma no com dos na por não as se é ao à das mais como foi* ontem *ser sua pelo pela mas já seu anos nos entre ou sobre aos ainda está* são dois *ter* tem ano *também* cento *às nas* há dia Lisboa até contos mil depois onde *mesmo* milhões foram vai este Portugal *muito* presidente três num *seus sem* numa quando esta vez hoje só segundo país durante contra agora apenas parte primeira duas pelos grande passado desde *todos* cerca era Governo será sido dias primeiro estão
ENPCport	*de a que e o um para não uma com do da se em os por as no na como é mais* era ele *à ao* disse *mas sua dos* eu estava *seu ou* tinha ela *das* quando me *ser* casa *nos muito ter* até isso bem *foi anos* lhe *já* minha *mesmo* meu vez tempo depois *seus* onde num *sobre sem* fazer só fora *pelo também* lá sempre você dizer tão nada *está pela* pessoas *ainda* Alice numa *todos* qualquer tudo lado coisa mãe dois dia suas pai havia dele vida *nas* antes outro nunca tinham outra agora Andrew (FALTA: entre, às e aos)

Natura/Minho	*de a e que o do da em para os com um dos no uma na as por se ao das à não é mais como ser sua foi* Braga *ou* ano *ainda pelo pela* dia *seu* contos *aos também já nos anos todos às entre* tem são este Câmara vai *nas* Natal *está sobre* ontem será pessoas *mas* pelas mil hoje até onde esta presidente Janeiro *ter mesmo* Municipal Minho cento há 1999 *sem* dois social projecto vida todas José *seus* suas trabalho *muito* Portugal foram forma disse jovens três pelos estão outros grande fazer concelho 1 vez Porto
ECI-EBR	*de a que e o do da não em um para com se é os uma no na as por mais como dos ao das à* eu *ou* me *sua seu* ele *ser mas nos foi* era *já mesmo sem* você *pelo pela* tem *está muito* quando só *também* até são *ainda* bem meu tudo *entre* ela *todos* isso *seus* vida vai *aos sobre* dois gente homem *nas* tempo minha pode dia nem onde sempre *anos* Brasil vez assim *às* há casa grande outro pra mundo aqui porque coisa nada num suas depois tão quem todo tinha agora *ter* lhe
ECI-EE	*de a e que do o da em os para dos as um uma das à no por* programa *na como com é* I&D *se* TI *ao* informação indústria investigação *ser mais* ESPRIT sistemas Comunidade desenvolvimento *não* suporte lógico *ou* projectos Comissão produção *sua anos sobre* resultados actividades *aos* tecnologias tecnologia será Estados trabalhos trabalho dados *pela* são nível industrial *entre* serão prazo áreas longo *todos* área objectivos *às nos* produtos mercado *seu* recursos projecto sistema Membros técnicas *pelo* objectivo gestão execução comunitária principais sector computador aplicação processo processamento *nas* modo grandes bases base acesso Europa Conselho parte necessário métodos (FALTAM 11)
MLCC-DEB	*de a que e o da do em os não para uma se um dos no é as com* Comissão *à na ao das por como sobre mais ser* Parlamento Presidente Conselho *nos* Comunidade Senhor *também mas pelo ou muito aos sua* deputado este relatório esta *às* países são *já* política *foi pela* Europeu *entre está seu* proposta isso questão *ainda todos* Europa tem parte há Europeia *ter* facto acordo fazer seja situação lugar bem pode neste *nas* forma *sem* trabalho agora debate qualquer vez só *mesmo* caso mercado alterações nossa medidas União dizer *seus* deve outros nome resolução propostas (FALTA: anos)
NILC/São Carlos	*de a o e que do da em para é no com um os na não uma dos por se as mais ao como* EDITORIA *das à foi ser pelo sua ou pela seu* São Paulo *está* tem *nos sobre entre mas* disse *também anos* ele *já* até Brasil ontem governo Folha dia vai ano diz *às* pode dois *ainda* presidente contra *nas muito* hoje *mesmo* Rio há *aos ter* só *seus* Segundo *sem* quando mil foram país três era estão será deve maior apenas Local pessoas isso Reportagem milhões tempo fazer EUA dias primeiro onde menos porque suas depois (FALTA: todos)
Comuns a todos os corpora menos 1	*a ainda anos ao aos as com como da das de do dos e em entre está foi já mais mas mesmo muito na nas no nos não o os ou para pela pelo por que se sem ser seu seus sobre sua também ter todos um uma à às é*

Tabela 3: As cem palavras mais frequentes em cada corpus

Seria de esperar, como aconteceu, que os corpora mais temáticos fossem os que se desviassem mais da média. Não é, pois, de estranhar que o anúncio do programa ESPRIT seja quase singular, enquanto que (as colectâneas) de textos literários tenham palavras menos características. De uma análise superficial das palavras únicas destas listas salienta-se claramente o carácter regional do Diário do Minho, assim como é bem patente o vocabulário do Parlamento Europeu.

Algumas particularidades do uso brasileiro (por oposição ao do uso português) são também facilmente identificáveis na lista das palavras mais frequentes: o aparecimento de *gente*, *pra* e *todo* em ECI-BR não é certamente devido ao acaso, nem a frequência de *ele* nos dois corpora brasileiros. Surpreendentemente, a palavra *há* pertence nos dois corpora brasileiros às cem mais frequentes, o que leva a concluir que é errada a suposição portuguesa de que se encontra definitivamente substituída por *tem* na norma brasileira (além, evidentemente, do seu uso como localizador temporal).

Os limites desta comparação são contudo aparentes, se atentarmos em que a palavra *Comissão* está tipicamente associada ao "europês". Contudo, o facto de termos dois corpora provindos da mesma realidade (MLCC-DEB e o ECI-EE) não permitiu que esse termo se salientasse, como devia, como característico desse discurso.

Não apareceram palavras inesperadas comuns a todos os corpora, com excepção, talvez, da palavra *anos*, o único nome, aliás. De facto, uma questão pertinente é a de ajuizar se a

avaliação das cem palavras mais frequentes não devia de facto excluir as palavras gramaticais. Em primeiro lugar, porque estas têm uma distribuição completamente diferente, como o demonstra Katz (1996). Em segundo lugar, por razões pragmáticas. Para a maior parte dos objectivos a que se subordinam análises lexicais (modelo da perplexidade num modelo de língua estatístico, recuperação da informação em textos, obtenção semi-automática de terminologia, lexicografia assistida por computador, etc.) são as palavras com conteúdo que interessam aos investigadores.

Obtivemos portanto as listas dos trinta nomes mais frequentes[3] em cada corpus, que apresentamos na Tabela 4.

Corpus	30 nomes mais frequentes, por ordem
Natura/Público	anos ano dia Lisboa contos Portugal presidente vez país parte Governo dias semana tempo Porto grupo Estado caso forma mercado Público ministro empresa pessoas países acordo fim situação cidade processo
ENPCport	casa anos vez tempo pessoas Alice lado coisa mãe dia pai vida Andrew porta vezes olhos parte mulher noite coisas cabeça Macon nome homem Oliver cima Harriet mão Stuart cidade
Natura/Minho	Braga ano dia contos anos Câmara Natal pessoas presidente Janeiro Minho projecto vida José trabalho Portugal forma jovens concelho Porto João parte euro cidade número moeda Maria Associação António actividades
ECI-EBR	vida gente homem tempo dia anos Brasil vez casa mundo coisa mulher povo lado meio trabalho vezes noite forma mão olhos nome coisas mãe cidade pai porta terra corpo senhor
ECI-EE	programa I&D TI informação indústria investigação ESPRIT sistemas Comunidade desenvolvimento suporte projectos Comissão produção anos resultados actividades tecnologias tecnologia Estados trabalhos trabalho dados nível prazo áreas área objectivos produtos mercado
MLCC-DEB	Comissão Parlamento Presidente Conselho Comunidade Senhor deputado relatório países política proposta questão Europa parte Estados-membros acto acordo situação lugar forma trabalho debate vez caso mercado alterações medidas União nome resolução
NILC/São Carlos	EDITORIA São Paulo Brasil governo Folha dia ano presidente Rio país Local pessoas Reportagem tempo EUA dias mercado Estado vez caso mês cidade empresa candidato preços mundo semana José empresas trabalho jogo Fernando

Tabela 4. Os trinta nomes mais frequentes em cada corpus

Convém referir que, ao contrário da maioria dos investigadores, não entrámos em conta com "palavras com espaços no meio", tais como *no entanto, a favor de* ou *papo seco*, nem com expressões fixas como *de ver a Deus* ou *tirar nabos da púcara*. As razões são puramente metodológicas e prendem-se com um desejo de replicabilidade e de simplicidade deste tipo de experiências, como foi magistralmente defendido em Kilgarriff (1997a). É nossa opinião, aliás, que a necessidade de selecção (e a própria selecção) de conjuntos de palavras gráficas como entidades individuais é dependente do objectivo ao qual se subordina a descrição da língua (Santos (1990) sugere um critério apropriado à tradução automática). Os leitores são portanto convidados a ter esta questão em mente na sua interpretação dos dados. Por exemplo, é provável que uma grande percentagem das ocorrências da palavra *de* se deva à sua participação em locuções gramaticais.

---

[3] Visto que os corpora não estão anotados gramaticalmente, foi preciso usar o conhecimento subjectivo da língua para retirar palavras como *este*, *pelos*, *cerca*, que embora também possam ser nomes, certamente se encontram na lista pelos seus usos como palavras gramaticais. Da mesma forma, aceitou-se *parte* apesar de também poder ser uma forma verbal e *jovens* apesar do adjectivo homónimo.

Em alguns casos foi, contudo, preciso analisar os resultados em pormenor. Por exemplo, a forma *final* com 3500 ocorrências no corpus Natura/Público apenas em 2200 casos correspondia a um nome, donde, refazendo as contagens, não se incluía já nos 30 nomes mais frequentes (de notar que além disso o seu uso nominal encontra-se equitativamente distribuído entre os casos de *o final* e *a final*).

Por outro lado, note-se que a própria definição de nome não é necessariamente consensual. Formas como *gente*, *número* ou *lado* (se maioritariamente presentes no contexto *a gente* ou na locução *por outro lado*) talvez não fossem incluídas como nomes por outros investigadores.

4.3. Riqueza de nomes próprios

Tentando aprofundar um pouco mais a análise dos corpora, criámos um detector automático de nomes próprios – assunto que tem vindo a ser muito debatido na literatura, veja-se p. ex. Mikheev et al. (1999) – baseado nos seguintes pressupostos:

(1) Palavras (ou sequências de palavras) com a primeira inicial maiúscula no meio da frase são nomes próprios. (Note-se que este critério inclui as iniciais.)
(2) Dois nomes próprios entremeados por *de do da dos* ou *das* são nomes próprios.
(3) Não se contabilizaram como nomes próprios aqueles ligados por *e*.
(4) Um nome próprio imediatamente a seguir à primeira palavra de uma frase inclui a primeira palavra dessa frase (excepto se esta pertencer a uma lista previamente especificada[4]); a preposição *de* e as suas contracções funcionam como no ponto precedente.

É evidente que não se pode esperar que um sistema baseado em tão pouca informação possa extrair precisamente todos e apenas os nomes próprios válidos nos corpora. Contudo, as várias opções foram pensadas para maximizar o número de nomes próprios detectados e minimizar o número de falsos positivos.

Assim, não se considerou o critério de nomes próprios ligados por *e* porque esse critério sobregerava claramente: Nos primeiros duzentos "nomes próprios" obtidos por essa regra apenas 58 poderiam ser considerados como tal, sendo 128 completamente inaceitáveis[5]. Comparando com o critério (2), note-se dos primeiros duzentos nomes próprios obtidos com a regra de *de* e suas contracções, apenas um caso foi considerado inaceitável e 9 suscitaram dúvidas.

Para avaliar a opção (4), observámos os primeiros duzentos nomes próprios incluindo a primeira palavra da frase, resultando em apenas 12 casos que não eram claramente nomes próprios.[6]

Os resultados aplicados a cada corpus produziram os resultados apresentados na Tabela 5. Esta tabela apresenta para cada célula, na primeira linha, o número de nomes próprios – com uma palavra, com duas, com duas incluindo *de*, etc. – e, na segunda, o número de nomes próprios distintos.

A percentagem de nomes próprios, para permitir uma comparação quantitativa entre os diversos corpora, é calculada dividindo o número absoluto de formas pertencentes à lista de nomes próprios pelo número total de formas no corpus. Não é possível obter facilmente um

---

[4] As palavras que constam da lista são: artigos, preposições: *de para por em com como segundo sob entre sobre à ao*, locuções com *de*: *além apesar através depois dentro atrás perto antes*, conjunções: *se desde mas quando porque mal enquanto que*, advérbios *entretanto talvez apenas só já embora hoje ontem amanhã*, demonstrativos *este aquele esse nesse neste*, modificadores do nome: *mesmo até também tanto nem*. Além disso, retiram-se automaticamente todas as palavras com clíticos.

[5] A grande maioria dos casos correspondia à conjunção de nomes próprios independentes, em alguns casos (muito raros) com elisão parcial, tal como os irmãos *Domingos e Dionísio Castro*, ou *os Presidentes Bill Clinton e Boris Ieltsin*.

[6] Três casos tinham um verbo em primeira posição, dois correspondiam a nomes no plural (*Imperativos do Mercado Interno, Jornalistas de Belgrado*), três correspondiam a nomes no singular (*Relatório do..., Prémio do..., Governo de...*), outros correspondiam a casos menos claros: *Ciclo Beethoven, Acordo Geral, Anos de Televisão* (mas note-se que estes casos são independentes de incluir a primeira palavra da frase ou não, representando os limites de um critério baseado na ortografia do autor do texto).

número semelhante em termos de tipos (nomes próprios distintos), visto que apenas podemos contar directamente o número de palavras isoladas. (O número de nomes próprios distintos é muito grande porque entra com as combinações de nomes próprios.) Apresentamos, contudo, a medida da percentagem de tipos de nomes próprios simples (só uma palavra) comparados com o número de tipos de palavras simples total.

Forma do nome próprio (palavras)	Natura/Público	ENPCport	Natura/Minho	ECI-EBR	ECI-EE	MLCC-DEB	NILC/São Carlos
1	118 181 16 430	4 691 907	51 717 6953	10 052 3 413	192 37	175 800 5 771	695 662 47 179
2	62 045 22 842	921 483	30 328 9 627	3 016 1 879	60 14	63 641 3 680	386 806 82 884
2+de	15 433 5 169	98 66	16 063 4996	826 532	7 6	10 795 1 202	60 885 14 403
3	6 450 3 671	83 55	5 344 2 759	421 349	3 3	2 941 1 039	50 944 20 336
3+de	6 023 2 959	24 22	10 026 4 665	295 239	2 13	3 005 693	293 807 10 870
4	693 542	1 1	1 067 712	46 43	2 1	198 142	6 624 3 359
4+de	1 122 742	0	3 721 2 103	77 72	0	1 264 176	6 535 3 143
>5 com ou sem *de*	546 410	7 6	2 947 2 116	49 46	2 1	146 88	2 855 1 700
% nomes próprios	8,79	3,72	12,7	4,81	3,07	6,38	9,74
% tipos simples	9,52	3,41	8,80	5,67	0,89	5,19	11,06

Tabela 5: A contabilização de nomes próprios em cada corpus

Verificamos que o texto jornalístico é o que possui maior profusão de nomes próprios (e inclui também aqueles que são mais complexos). Ainda que, em valor absoluto, os debates do Parlamento Europeu contenham o maior número de nomes próprios simples, a variedade é muito menor, como se vê pela contabilização dos tipos.

Também seria interessante comparar, para ir mais fundo nestas observações, eram os nomes próprios mais frequentes em cada corpus e aqueles que eram comuns a todos. Reproduzimos na Tabela 6 os quinze nomes próprios simples[7] mais frequentes em cada corpus.

Corpus	15 nomes próprios simples mais frequentes, por ordem
Natura/Público	Lisboa Portugal Governo Porto Estado Público Câmara José Abril Presidente João António Silva República Comissão
ENPCport	Alice Andrew Macon Oliver Harriet Stuart Celia Rainha David Utz Rembrandt Paul Aristóteles Sarah Gillian

[7] Mais uma vez, nem todos os elementos da lista são "nomes próprios", mas sim elementos que são considerados pelo programa de sua detecção. *Nacional* faz certamente parte de vários nomes próprios, e daí a sua frequência elevada, enquanto que *Sr.* ou *D.* são títulos (abreviatura de Senhor e de Dom) que também não são nomes em si (mas constituem parte integrante de nomes próprios).

Natura/Minho	Braga Câmara Natal S. Janeiro Municipal Minho José D. Portugal Porto João Maria Associação António
ECI-BR	Brasil Rio Governo Estado João Maria Presidente José Sr. Pedro Estados Nacional D. República Carlos
ECI-EE	I&D TI ESPRIT Comunidade Comissão Estados Membros Europa Conselho Informação Unidos Comité COM Tecnologias CIMLE
MLCC-DEB	Comissão Parlamento Presidente Conselho Comunidade Senhor Europa União Tratado Senhora Grupo Estados Maastricht Doc. Assuntos
NILC/São Carlos	EDITORIA São Paulo Brasil Folha Rio Local Reportagem EUA Estado FHC José COTIDIANO Fernando Lula

Tabela 6: Os quinze nomes próprios simples mais frequentes de cada corpus

A leitura desta tabela já nos permite ter uma ideia mais precisa sobre o conteúdo dos textos que compõem cada corpus, ainda que mais uma vez não seja trivial escolher as palavras que melhor caracterizam uma colecção (Kilgarriff, 1996).

4.4. Verbos de percepção

A distribuição de alguns verbos também pode ser interessante. Em Santos (1998) são discutidas algumas diferenças entre verbos de percepção do português e do inglês, notando que o verbo *imaginar* tem uma distribuição sintáctica semelhante aos verbos de percepção no que se refere às construções destes verbos com um sintagma verbal objecto. Como observação preliminar, medimos o número de ocorrências destes verbos em cada corpus, simplesmente adicionando o número de ocorrências de cada forma (sem contar com os particípios passados).[8]

Corpus	*ver*	*ouvir*	*sentir*	*imaginar*	Total
Natura/Público	3 859	1 020	979	257	6 115 (0,98)
ENPCport	425	170	280	90	965 (4,16)
Natura/Minho	90	34	50	8	182 (0,64)
ECI-EBR	1 078	402	440	120	2 040 (2,86)
ECI-EE	2	0	0	0	2 (0,08)
MLCC-DEB	6 287	1 739	1 637	292	9 955 (1,12)
NILC/São Carlos	17 723[9]	4 246	5 778	2 217	29964 (0,95)

Tabela 7: A distribuição de quatro verbos em cada corpus

É interessante notar que a proporção relativa de *ver* é semelhante em todos os corpora, enquanto que predominância relativa de *sentir* e *ouvir* se altera, a favor de *sentir*, no discurso literário. O uso destas palavras (a sua ocorrência por milhar de palavras de cada corpus, indicada na última coluna entre parênteses) é também inequivocamente mais frequente neste tipo de discurso. Por outro lado, é curioso que os debates do Parlamento Europeu contenham mais lexemas de percepção ou imaginação do que o texto jornalístico, seja ele português ou brasileiro.

4.5. *Onde* vs. *aonde*; e *quando*?

O estudo de algumas construções pode ser também revelador de diferenças entre os corpora. Escolhemos as palavras *onde*, *aonde*, *donde* e *quando* por duas razões: Em primeiro lugar, o

[8] Foi preciso, naturalmente, investigar manualmente todas as formas homónimas com outras palavras, tais como *vimos*, *verão*, *sente*, *via*, *vendo*, etc.
[9] Este valor foi parcialmente estimado.

nosso interesse na variação *onde/aonde* (veja-se Santos, em prep.); em segundo, porque estas palavras poderão ser indicadores de localização espacial e temporal.

Corpus	*onde*	*aonde*	*donde*	Total **onde*	*quando*
Natura/Público	6 327	15	32	6 374	6 097
ENPCport	338	2	7	347	731
Natura/Minho	274	0	2	276	207
ECI-EBR	707	5	5	717	1 311
ECI-EE	1	0	1	2	18
MLCC-DEB	3 354	18	148	3 520	8 167
NILC/São Carlos	20 793	137	141	21 071	35 750

Tabela 8: A distribuição de quatro "localizadores" em cada corpus

Ainda que, evidentemente, a presença destes itens lexicais não seja uma medida fiel da atenção prestada ao espaço e ao tempo, não deixa de ser interessante reparar que o texto jornalístico liga muito mais ao lugar do que os outros tipos de texto, em que a proporção "natural" *quando*/*onde* é de 2 para 1.

4.6. Comentários finais

Com trabalho futuro, é necessário avançar para uma caracterização estatística dos corpora, das suas diferenças e grau de semelhança à imagem de Kilgarriff & Rose (1999) e de Katz (1996), assim como é necessário efectuar uma caracterização da língua portuguesa com base em corpora, replicando por exemplo os estudos de Medeiros, Marques, & Santos (1993) sobre corpora muito maiores.

Contudo, pensamos que o trabalho descrito no presente artigo é necessário como desbravamento inicial. Os dados que se extraíram, embora extremamente superficiais e porventura pouco informativos, podem fornecer algumas pistas sobre o caminho que é preciso percorrer, assim como ilustram a informação fácil de obter e também as deficiências a ela associadas.

Talvez o maior contributo do presente artigo consista na ilustração da dificuldade de quantificar diferenças entre géneros de uma língua sem um trabalho profundo de preparação dos materiais e sem uma investigação aturada dos métodos da área de linguística com corpora. De facto, ainda que para teste do desempenho e cobertura de sistemas de processamento de linguagem natural não haja dúvidas sobre a necessidade de corpora de texto, o uso destes para a investigação em linguística ainda não conseguiu refutar, na minha opinião, o que poderíamos chamar "regra de Franckel": "O grau de estabilidade [de um fenómeno linguístico] é inversamente proporcional à imediatez da sua observação" (Franckel, 1989:24, tradução minha).

5. Agradecimentos

Agradeço os comentários e sugestões pertinentes de Signe Oksefjell e Elisabete Ranchhod.

Estou, além disso, muito grata aos compiladores dos corpora utilizados neste trabalho, evidentemente impossível sem eles. Agradeço, pois, especialmente, a José João Dias de Almeida, Stig Johansson, Myriam Ramsey, Henry S. Thompson e Osvaldo Oliveira pela presteza com que forneceram os corpora e a informação necessária.

6. Referências

Armstrong, Susan, Kempen, Masja, McKelvie, David, Petitpierre, Dominique, Rapp, Reinhard, & Thompson, Henry S. (1997). Multilingual Corpora for Cooperation - Revised final report. Disponível na Internet no endereço www.ltg.ed.ac.uk/~dmck/mlcc.ps.

Armstrong, Susan, Kempen, Masja, McKelvie, David, Petitpierre, Dominique, Rapp, Reinhard, & Thompson, Henry S. (1998). Multilingual Corpora for Cooperation. In Rubio, Antonio, Gallardo, Natividad, Castro, Rosa, & Tejada, Antonio (Orgs.), *Proceedings of The First International Conference on Language Resources and Evaluation* (Granada, 28-30 May 1998) (Vol.2, pp.975-80).

Christ, Oliver, Schulze, Bruno M., Hofmann, Anja, & Koenig, Esther (1999). The IMS Corpus Workbench: Corpus Query Processor (CQP): User's manual. Institute for Natural Language Processing, University of Stuttgart, March 8, 1999 (CQP V2.2). Disponível na Internet no endereço http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/.

Dale, Robert (1997). A Framework for Complex Tokenisation and its Application to Newspaper Text. In *Proceedings of the Second Australian Document Computing Symposium* (Melbourne, Australia, April 5, 1997).

Franckel, Jean-Jacques (1989). *Etude de Quelques Marqueurs Aspectuels du Français*. Genève: Librairie Droz.

Johansson, Stig, Ebeling, Jarle, & Hofland, Knut. (1996). Coding and aligning the English-Norwegian Parallel Corpus. In Aijmer, K., Altenberg , B., & Johansson, M. (Orgs.), *Languages in Contrast* (pp. 87-112). Lund: Lund University Press.

Johansson, Stig, Ebeling, Jarle & Oksefjell, Signe (1999). English-Norwegian Parallel Corpus: Manual. Oslo: Department of British and American Studies, University of Oslo. Disponível na Internet no endereço www.hf.uio.no/iba/prosjekt/ENPCmanual.html.

Katz, Slava M. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering, 2* (1), 15-59.

Kilgarriff, Adam (1996). Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition* (Sussex, April 1996) (pp 33-40).

Kilgarriff, Adam (1997a). Putting frequencies in the dictionary. *International Journal of Lexicography, 10* (2), 135-155.

Kilgarriff, Adam (1997b). Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proc. 5th ACL Workshop on Very Large Corpora*, Beijing and Hong Kong. Também Report ITRI-97-07, Information Technology Research Institute, University of Brighton.

Kilgarriff, Adam & Rose, Tony (1999). Measures for corpus similarity and homogeneity. In *Proc. 3rd Conf. on Empirical Methods in Natural Language Processing* (Granada, June 1998) (pp 46-52). Também Report ITRI-98-07, Information Technology Research Institute, University of Brighton.

McKelvie, D. & Thompson, H. S. (1994). TEI-Conformant structural markup of a trilingual parallel corpus in the ECI Multilingual Corpus 1. In *Proceedings of the 2nd Annual Workshop on Very Large Corpora - WVLC2* (Kyoto, 4 August 1994) (pp. 7-18).

Medeiros, José Carlos, Marques, Rui, & Santos, Diana (1993). Português quantitativo. In *Actas do 1.o Encontro de Processamento de Língua Portuguesa (Escrita e Falada) - EPLP'93* (Lisboa, 25-26 Fevereiro 1993) (pp.33-8).

Mikheev, Andrei, Moens, Marc, & Grover, Claire (1999). Named entity recognition without

gazetteers. In *Proceedings of EACL'99* (Bergen, 8-12 June 1999) (pp.1-8). Bergen: ACL.

Nunes, M.G.V., Vieira, F.M.C., Zavaglia, C., Sossolote, C.R.C., & Hernandez, J. (1996a.) A construção de um léxico para o português do Brasil: lições aprendidas e perspectivas. In *Proceedings of the II Workshop on Computational Processing of Written and Spoken Portuguese* (Curitiba, 23 a 25/10/96) (pp. 61-70). Disponível na Internet no endereço www.icmc.sc.NILC.br/~mdgvnune/download/curitilex.ps.gz

Nunes, M.G.V., Turine, M.A.S., Martins, R.T., Ghiraldelo, C.M., Oliveira, M.C.F., Montilha, G., Hasegawa, R., & Oliveira Jr., O.N. (1996b). Desenvolvimento de um sistema de revisão gramatical automática para o português do Brasil. In *Proceedings of the II Workshop on Computational Processing of Written and Spoken Portuguese* (Curitiba, 21 a 22/10/96) (pp. 71-80). Disponível na Internet no endereço www.icmc.sc.NILC.br/~mdgvnune/download/curitiregra.ps.gz

Oksefjell, Signe (1999). ENPC: Um corpus paralelo que inclui o português. In Marrafa, Palmira, & Mota, Maria Antónia (Orgs.), *Actas do I Workshop sobre Linguística Computacional da Associação Portuguesa de Linguística* (Lisboa, 25-27 de Maio de 1998). Lisboa: APL.

Oksefjell, Signe (no prelo). A description of the English-Norwegian Parallel Corpus: Compilation and further developments. *International Journal of Corpus Linguistics*.

Oksefjell, Signe & Santos, Diana (1998). Breve panorâmica dos recursos de português mencionados na Web. In Lima, Vera Lúcia Strube de (Org.), *Anais do Terceiro Encontro de Processamento da Língua Portuguesa (Escrita e falada) - PROPOR'98* (Porto Alegre, 3-4 novembro 1998) (pp.38-47).

Peters, Wim, Cunningham, Hamish, McCauley, Clare, Bontcheva, Kalina, & Wilks, Yorick (1998). Uniform language resource access and distribution. In Rubio, Antonio, Gallardo, Natividad, Castro, Rosa, & Tejada, Antonio (Orgs.), *Proceedings of The First International Conference on Language Resources and Evaluation* (Granada, 28-30 May 1998), (Vol. 1, pp.13-7).

Santos, Diana (1990). Lexical gaps and idioms in Machine Translation. In Karlgren, Hans (Org.), *Proceedings of COLING'90* (Helsinki, August 1990) (Vol 2, pp.330-5).

Santos, Diana (1998). Perception verbs in English and Portuguese. In Johansson, Stig & Oksefjell, Signe (Orgs.), *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies* (pp.319-342). Amsterdam: Rodopi.

Santos, Diana. (Em prep.). Aonde vamos em relação a *aonde*. Em preparação.

Santos, Diana & Ranchhod, Elisabete (1999). Ambientes de processamento de corpora em português: Comparação entre dois sistemas. In *Actas do IV Encontro sobre o Processamento Computacional da Língua Portuguesa (Escrita e Falada)* (Évora, 20-21 de Setembro 1999).

Thompson, H., Armstrong-Warwick, S., McKelvie, D., et al. (1994). Data in your language: The ECI Multilingual Corpus 1. In *Proceedings of the International Workshop on Shareable Natural Language Resources*. Nara.

Wittmann, Luzia, Pêgo, Tânia & Santos, Diana (1995). Português do Brasil e de Portugal: alguns contrastes. In *Actas do XI Encontro da Associação Portuguesa de Linguística* (Lisboa, 2-4 de Outubro de 1995) (pp.465-487). Lisboa: APL.