Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Corpus Nilc:
descrição e análise crítica com vistas ao projeto Lacio-Web

Gisele Montilha Pinheiro

Sandra Maria Aluísio

NILC-TR-03-03

Fevereiro de 2003

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Agradecimentos

Um agradecimento especial ao *Jorge Augusto Teles* pela colaboração no trabalho de construção e formatação de gráficos e tabelas, bem como da formatação do texto. E ao *Ricardo Hasegawa* pelo desenvolvimento do programa de contagem de freqüência e pelas discussões sobre as estatísticas de desempenho do ReGra.

Índice

I. Introdução

Este relatório foi construído com o intuito de fornecer as mais variadas informações sobre o Corpus Nilc, em 2002, tendo em vista que tais apontamentos serão norteadores para o trabalho de construção do *Projeto Lacio-Web*, em curso no Nilc.

A fim de esclarecermos determinadas decisões envolvendo o design do Corpus Nilc, iniciamos este documento com o relato das principais etapas pelas quais o Corpus passou nesses 10 anos de sua existência. A seguir, na Seção III, dedicamo-nos a um trabalho essencialmente descritivo, que é o resultado da análise do Corpus Nilc – situação em setembro de 2002.

Nossa análise, como dissemos, tem estreita relação com a construção de um novo corpus, mais especificamente o *Lacio-Ref* – o corpus de Referência do Lacio-Web. Em virtude disso, a descrição de que tratamos na Seção III inclui apenas o corpus corrigido e um diretório do semicorrigido (o corpus universitário). Esse recorte teve como parâmetro o fato de serem textos com grandes chances de revelarem o uso do português culto e, portanto, serem de potencial aproveitamento no Lacio-Ref.

Na Seção IV, apresentamos uma breve reflexão sobre um dos principais aspectos decorrentes da existência do Corpus Nilc: a contagem de palavras. Caracterizando o Corpus em termos numéricos, trazemos dados curiosos que podem demonstrar de que forma o design de corpus afeta as análises lingüísticas. Trata-se de uma discussão superficial, mas ao mesmo tempo provocadora, na tentativa de confirmar, através de dados concretos, a necessidade de reformulação do Corpus Nilc.

Por último, na Seção V, definimos os procedimentos mais relevantes para a construção do Lacio-Ref, a partir daquilo que detectamos na análise do estado atual do Corpus Nilc.

Neste relatório, constam ainda: a) um sumário, no final da Seção II, apresentando, resumidamente, os resultados da análise do Corpus; e b) três apêndices: o primeiro, em que apresentamos a estrutura do Corpus, parte que foi investigada neste relatório; o segundo, onde reunimos os gráficos que ilustram o tamanho dos corpora examinados; e o terceiro, que se constitui num extrato da tabela de freqüência das ocorrências no Corpus.

II. Apresentação histórica: uma contextualização

Em meados de 1993, com o convênio firmado entre a USP e a ITAUTEC/PHILCO, teve início o projeto de desenvolvimento do *ReGra* – programa de revisão gramatical para auxílio na redação de textos escritos em português do Brasil. Por esta ocasião, deu-se início a construção de diversas ferramentas computacionais, visando a dar suporte à tarefa de revisão textual; dentre esse material de suporte, foi elaborado um banco de textos o qual, mais tarde, se tornaria o Corpus de base para diversos aplicativos desenvolvidos pelo Nilc. Isso é, pois, o que chamamos de **Corpus-Nilc**, ou simplesmente, CN.

Conforme esclarece Nunes et al. (1995, cap. 5.2), a construção do CN teve por objetivo: a) atender à necessidade de efetuar os testes da ferramenta de revisão com textos reais, e b) fornecer a base empírica do uso de algumas formas gramaticais do português. Como veremos, essas metas viriam, mais tarde, a influenciar decisivamente o perfil do CN como um todo.

Para cumprir o item (a), a seleção dos textos do Corpus foi determinada, em grande parte, pela necessidade da verificação dos chamados "falsos erros"[1] evidenciados num corpus constituído por "textos escritos por autores experientes, revisados e corrigidos" (op.cit., p. 12). Já os casos de "omissão", por esse raciocínio, seriam testados nos demais conjuntos, com especial ênfase na base de textos não-corrigidos[2]. O CN, a partir disso, tornou-se uma base tripartida, compondo-se de um conjunto de textos corrigidos, de textos semicorrigidos e de não-corrigidos (veja mais na seção II).

Essas três categorias de corpus, então, passaram a comportar aqueles textos que fossem selecionados de acordo com um critério especificamente justificado pela tarefa computacional a que se prestava, a saber, a revisão ortográfico-gramatical. Desse modo, por exemplo, considerou-se como texto corrigido um jornal ou um livro didático, supondo, para isso, que a fim de atingirem publicação, os textos passaram por diversas etapas de revisão e que, salvo alguma percentagem pouco significativa, eles seriam textos livres de inadequações lingüísticas, i.e., livre de erros.

Contudo, para além dessa influência motivada pelo desempenho do processamento da ferramenta, o CN também viria a ser determinado pelo item (b) indicado atrás. Por certo, o trabalho de investigação lingüística com vistas à correção sempre se pautou na gramática normativa (português culto). No entanto, não se prescindiram das ocorrências não-registradas nas gramáticas tradicionais e

[1] Para uma discussão sobre o emprego da nomenclatura que ora utilizamos, como falsos erros/falsos negativos, omissão/verdadeiros negativos, etc., cf. Nunes et. al *Relatório dos Testes Comparativos entre Diferentes Versões do Revisor Gramatical ReGra*. (NILC-TR-00-8). Junho 2000, 08p., disponível http://www.nilc.icmc.usp.br/nilc/publications.htm#ConferencePapers.

nos dicionários padrões, ou seja, foi necessário contar com um corpus para efetivar análises lingüísticas.

Nesse sentido, desde cedo o CN precisou abranger um conjunto exaustivo de textos, ampliando, o mais possível, os registros de escrita comum. A compilação desse material, no entanto, orientou-se pelo critério da facilidade na aquisição dos textos, o que terminou gerando grupos muito extensos de textos ao lado de conjuntos quantitativamente muito reduzidos. Este é o caso do corpus jornalístico que, na época de construção, teve garantida a exportação de um cd-rom inteiro, contendo todos os textos de 1994 do jornal *Folha de São Paulo*. Para o CN (situação em 2002) isso significou um conjunto de 3.342 arquivos (75,8% do corpus corrigido). Outro exemplo representativo da exaustividade do CN em determinados gêneros e tipos de textos são as redações de vestibulandos, do corpus não-corrigido, resultante do acordo entre o Nilc e a Fuvest, que forneceu todas as redações produzidas no vestibular de 1994 pelos seus candidatos. Como conseqüência, a pasta relativa às redações soma 2.299 textos, representando 94,3% do corpus não-corrigido.

Na medida em que o CN tomava corpo e respondia satisfatoriamente aos propósitos de suporte do desenvolvimento do ReGra, seu prestígio dentro da comunidade de PLN em português crescia entusiasticamente. Muitos foram os pedidos de utilização do Corpus, argumentando-se que um corpus como o que tinha Nilc não existia para o português do Brasil e que, para diversas pesquisas em desenvolvimento na época, era o caso de se tomar um corpus já pronto[3]. Disso resultou a motivação para se ampliar o CN, ainda que a disponibilidade dele para outros pesquisadores constituísse um problema, já que para muitos textos o Nilc não tinha a concessão dos direitos autorais.

A expansão do CN era um dos pontos que, desde 1995, os próprios projetistas indicavam como perspectiva de trabalho futuro. Ao lado dela, a organização dos textos e o balanceamento do corpus surgiam como preocupação. Vejam-se, a esse respeito, as seguintes passagens de Nunes et. al (1995):

> (1) "A organização atual [do Corpus], porém, ainda não é definitiva, pois na medida em que o tratamento computacional da linguagem se sofistica e se especializa, será necessário também modificar o corpus a fim de adequá-lo às exigências dos testes das ferramentas." (op. cit., p. 13)
>
> (2) "O corpus ainda não conta com textos literários (...). É de grande importância a incorporação, num futuro próximo, de tais textos ao corpus, pois neles podem ocorrer construções lingüísticas singulares. Também num futuro próximo, pretende-se incorporar transcrições de textos orais (...). Além de textos

[2] A respeito dos dados extraídos desse trabalho de tabulação e estatísticas do desempenho do ReGra, veja-se trabalho recente (Hasegawa, R.; Martins, R.T.; Nunes, MG.V. *ReGra 2002: Características e Desempenho*. NILC-TR-02-8, ICMC-USP, Junho 2002, 14p.) disponível em http://www.nilc.icmc.usp.br/nilc/publications.htm#ConferencePapers.

[3] Falamos dos anos 1993-96, quando o corpus construído pela equipe do Prof. Borba (UNESP/Araraquara) ainda era bastante incipiente e não disponível para outros pesquisadores. Outros corpora que hoje são referências no âmbito do português do Brasil não existiam, como é o caso do CETEN-Folha e do CRPC.

> dessas duas classes, pretende-se também incorporar textos de outras áreas do conhecimento, como por exemplo, culinária, humor, ficção científica, entre outras." (idem, ibidem)
>
> (3) "Dada a evolução da pesquisa, será também necessária a elaboração de um balanceamento entre os tipos de textos para levantamentos mais confiáveis de freqüência de palavras – material de auxílio ao léxico; e de expressões e/ou estruturas sintáticas- auxílio às regras gramaticais." (idem, ibidem)
>
> (4) "Pretende-se também reorganizar o corpus, baseando-se em estudos sobre tipologia de textos que propõem classificações a partir de características intrínsecas a eles." (idem, p. 13/14)

Em suma, as declarações acima demonstram que os primeiro envolvidos na construção do CN tinham uma percepção bastante nítida das restrições do Corpus. E assumiam que as escolhas já realizadas para a organização e classificação dos textos eram decisões guiadas pela força da aplicação (a revisão automática) e do desenvolvimento de módulos de apoio (o léxico, por exemplo). Assim, ao largo das ambições que se projetavam, como nos aponta a passagem (2), com o passar dos anos de pesquisa a avaliação presente em (1) demonstrou-se sempre atual.

Sobre o balanceamento, do qual nos fala a passagem (3), notamos que o desequilíbrio entre tipos e gêneros de textos no CN já era evidente àquela época. Já a passagem (4) aponta para a necessidade de uma reformulação do Corpus, buscando-se pelo refinamento através do estudo de tipologia de texto. Mas, efetivamente, o que se deu no decorrer dos anos foi tão somente a expansão do Corpus, empreendida paulatinamente e sofrendo com a escassez de aparato técnico na fundamentação de suas escolhas.

As expectativas das ferramentas desenvolvidas pelo grupo e questões relacionadas à política de financiamentos tornaram especiais as condições de ampliação do Corpus. Surgiram especificidades, tais como a inserção de um conjunto de textos versando apenas sobre o tema futebol – fruto de um trabalho desenvolvido, em 1998, no âmbito do projeto UNL sobre tradução automática[4]. E, como conseqüência dessa falta de controle e critérios na ampliação do corpus, atualmente essa pasta de nome "Esportes" (corpus jornalístico) encontra-se no alvo das exclusões, quando a análise lingüística requer um corpus lexicalmente abrangente.

Em 2000 surge um Relatório Técnico em que se reportam algumas alterações do Corpus envolvendo a expansão. Registrando a direção que o desenvolvimento do Corpus tomou no decorrer desse período (1995 a 2000), o relatório indica que

[4] A respeito do projeto UNL, veja-se Oliveira Jr. et. al *O uso de interlíngua para comunicação via Internet: O Projeto UNL/Brasil*. Série de Relatórios do NILC. NILC-TR-01-3, Julho 2001, 14p., disponível em http://www.nilc.icmc.usp.br/nilc/publications.htm#ConferencePapers.

> "categorias mais abrangentes foram criadas para a classificação dos arquivos, principalmente na porção corrigida do Corpus. (...) Além disso, após uma revisão geral, foram feitas outras modificações: o diretório 'correspondência' foi reclassificado, sendo movido da parte dos textos corrigidos para a parte dos não-corrigidos; retirou-se do corpus todos os arquivos com textos em inglês; e foram retirados alguns arquivos com textos repetidos." (Kuhn, 2000, p. 1)

As reformulações apontadas acima, apesar de representarem melhorias no Corpus, foram alcançadas a partir da intuição (ainda que de um especialista) e do material a ser manipulado. Disso resultou uma apresentação mais refinada das pastas, sem que, no entanto, significasse uma alteração expressiva do perfil original do CN.

Mais significativas que as reformulações, então, foram as inserções promovidas nesse período. Conforme o relatório esclarece:

> "O trabalho principal que se realizou foi o aumento considerável dos textos da categoria Literários (209 arquivos novos), além de incluir-se arquivos novos de história, textos jurídicos, textos de enciclopédia, textos de revistas e jornais, entrevistas, relatórios e teses. O número total de palavras incluídas é 2.375.130." (idem, ibidem)

A expansão do CN, nas direções apontadas acima, parece ter focalizado apenas o aspecto quantitativo. Ainda que isso tenha contribuído para construção da representatividade de algumas categorias de textos, sobretudo aquelas antes não contempladas, podemos perceber que a organização textual, o controle da inserção de textos, a catalogação, etc., continuaram a constituir-se planos para um trabalho futuro com o CN.

Então, em 2001, um novo projeto de corpus teve vez no Nilc. A proposta do Projeto Lacio-Web veio contemplar as condições requeridas para o CN desde 1995 e, mais que isso, apresentar ao público um corpus com tratamento refinado de design, de organização textual e de representatividade. Espera-se, portanto, que o know-how adquirido com o CN e o próprio conteúdo dele somem vantagens para o desenvolvimento dos corpora que se projeta para o Lacio-Web.

III. Descrição interna do CN

No Anexo I, podemos observar a estrutura interna do CN e verificar que ele se compõe de:

a) três diretórios macros: os textos corrigidos, os semicorrigidos e os não-corrigidos;

b) diretórios micros em cada um dos conjuntos de (a), diferenciados pela especificidade de gênero textual. Por isso, no corpus corrigido, existem os diretórios jornalístico, jurídico, didático, etc.; no semicorrigido, os diretórios institucionais, universitários, etc.; e no não-corrigido, os diretórios orais, redações, etc.;

c) subdiretórios, especificando (b), que caracterizam fontes diferentes dos textos, ou a diferença segundo o tipo de texto, ou mesmo o assunto (ex.: diferença de tipo de texto: correspondência vs. redação, em corpus não-corrigido; diferença de assunto: esportes vs. futebol, em corpus corrigido; diferença de fonte: universitários vs. entrevistas, em corpus semicorrigido);

d) pastas, alocando categorizações mais específicas dos textos de cada subdiretório (ex.: a pasta 1colegio, do subdiretório didáticos em geral do corpus corrigido);

e) arquivos, que representam o texto propriamente dito.

O CN se compõe de um corpus do português do Brasil, no seu registro escrito, subdividindo os textos nos três diretórios macros. Além do que dissemos anteriormente sobre a influência da revisão textual para a publicação, esta característica da subdivisão se deve ao fato de que os textos foram selecionados de acordo com o tipo de receptores do material. Tal é o que esclarece com Nunes et al. (1995).

> "Os textos estão agrupados em três classes:
>
> a) textos publicados para grande número de leitores (...);
>
> b) textos publicados para pequeno número de leitores ou não publicados (...);
>
> c) textos não corrigidos, escritos por pessoas de nível médio de escolaridade (2º grau completo) e universitários" (p. 12).

O corpus corrigido (Cc) se constitui de 5 diretórios macros. E, a partir daqui, são esses conjuntos que passamos a descrever em maiores detalhes.

III.1 Corpus Jurídico de Cc

Este conjunto de **75** arquivos agrupa:

- o arquivo "Constituição" – texto de 1988, em sua forma integral;

- o subdiretório “Jurídicos em geral”, que diferencia os arquivos segundo os tipos distintos de texto;

No subdiretório, os arquivos não registram os dados tradicionais de referência (autor, publicação, data, etc.), mas são textos de fácil recuperação dessas informações. Os decretos, as leis e portarias, por exemplo, foram extraídos do DOU (Diário Oficial da União) e do CPC (Conselho da Ciência e Tecnologia).

Essa questão da fonte de importação dos textos, por sua vez, influenciou decisivamente o subdiretório. Uma vez que havia facilidade de acesso, os textos adquiridos versaram exclusivamente sobre temas ligados ao campo da informática. Não há, portanto, equilíbrio de assuntos nos arquivos das Leis, dos Decretos e das Portarias, o que deve ser reparado no novo corpus.

Quanto aos Códigos, que também constituem esse subdiretório, são textos que se apresentam como arquivos no corpus jurídico, o que vem a ser um equívoco de organização. Uma vez que existem subcategorias de códigos (código de processo e código anotado), os textos do código deveriam compor um diretório específico contendo as subcategorias individualizadas.

Os códigos compilados no CN são os chamados “códigos anotados”, i.e., aqueles que apresentam comentários especializados ao final de cada dispositivo. Além disso, os dois códigos existentes são “de processo”, outra categoria específica de documento jurídico. Por isso, esses textos deveriam ser classificados separadamente: num diretório, códigos vs. códigos de processo, e em outro, códigos anotados.

É mister salientar que, para efeito de documentação, os códigos, especialmente, devem estar devidamente acompanhados do Termo de autorização de uso, sobretudo os anotados, já que há uma autoria específica (o jurista que comenta) na produção do texto como um todo.

Quanto aos dados estatísticos, o Anexo II, onde tabulamos a contagem de palavras, mostra que o corpus Jurídico há grande volume de dados concentrado nos arquivos dos Códigos. No entanto, na comparação entre o número de palavras existentes e as palavras repetidas, vemos que arquivos menores, tais como a Constituição e as Portarias, são também importantes, porquanto apresentem menos repetição de termos no conjunto total dos textos. O corpus jurídico, assim, carece de melhor organização, já que os textos inseridos parecem representar satisfatoriamente o campo de conhecimento.

Dito isso, não há mais considerações relevantes a serem feitas com relação ao corpus jurídico.

III.2 Corpus Didático de Cc

Este conjunto de **66** arquivos agrupa:

- dois arquivos gerais: “História” e “Enciclopédia”;
- o subdiretório “Didáticos em geral”, subdividido em séries escolares determinadas;

Os arquivos gerais de que se fala são conjuntos de textos extraídos de CD-ROMs: o “História” vem do CD ___ e o “Enciclopédia”, do CD ___. Ambos os materiais foram adquiridos pelo Nilc com o propósito da ampliação do CN. Na época, o Corpus necessitava de textos versando sobre assuntos diversificados (culinária e religião, por exemplo), a fim de subsidiar o Léxico do ReGra com tipos específicos de ocorrências, tais como os nomes próprios (neste caso, a Enciclopédia fornecia textos com nomes de cidades, pessoas, estradas, etc.), e termos vinculados a campos específicos do conhecimento (à botânica e à geografia, por exemplo).

Mas, apesar dessas vantagens significativas, os dois arquivos não parecem inspirar confiança quanto à autenticidade da autoria. Boa parte dos dados de referência da enciclopédia História não está disponível no CD-ROM de onde partiu a compilação, o que dificulta, sobremaneira, a catalogação dos textos.

De maneira geral, a extração dos dados de referência nos dois arquivos enciclopédicos é prejudicada pela ausência dessas indicações no conjunto das obras e nos textos propriamente ditos[5]. Por isso, como se garante que os textos dessas enciclopédias são autênticos em vez de compilados, ainda que parcialmente, de outras fontes (i.e., como garantir que os autores são os mesmos que os organizadores dos compêndios)? No caso do Lacio-Web, que precisa disponibilizar ao usuário essas informações de cabeçalho, o problema da autenticidade inviabiliza a presença do corpus enciclopédico no corpus, a menos que essa lacuna seja preenchida pelo diálogo com os editores do material.

Ainda com respeito à presença do corpus enciclopédico, importa ressaltar que, muito embora a diversidade temática tenha colaborado para os propósitos do CN, os textos não estão classificados em grupos temáticos e estão, além disso, agrupados em arquivos gigantescos, inviabilizando uma consulta por assunto, por exemplo. Um tratamento peculiar deve ser promovido neste material, a fim de etiquetar cada um dos textos de acordo com o assunto e o campo de conhecimento. Nesse caso, uma organização textual satisfatória seria a apresentação individual dos textos, os quais, mesmo quando pequenos, constituiriam arquivos separados de pastas temáticas.

A importância da diversidade de assunto oferecido pelas enciclopédias, por sua vez, pode ser observada no quadro do Anexo II. Nele constatamos que, a despeito do volume de dados do

[5] Os dados bibliográficos de um texto escrito são itens que compõem o chamado “cabeçalho” dos textos num corpus. Hoje existem técnicas computacionais que se dedicam a sofisticar o design de cabeçalhos de corpus, sobretudo pela importância que ele tem para a construção de sistemas automáticos de busca e consulta ao corpus. No Lacio-Web fazemos uma reflexão sobre esse aspecto do design do Corpus, onde aplicamos as orientações do TEC (*Translation English Corpus*). Cf. www2.umist.ac.uk/ctis/research/TEC/tec_home_page.htm

subdiretório “Colégio” (Didáticos em geral), é nas enciclopédias que a pluralidade terminológica está representada (vide quadro comparativo). Mas esse dado, por sua vez, encontra-se viciado devido à especificidade da constituição do subdiretório Colégio.

Na verdade, “Didáticos em geral” se compõe de textos extraídos de livros didáticos utilizados na rede estadual de ensino à época da compilação (1995). Ocorre que, em virtude de algumas particularidades das disciplinas, não foram aproveitados: a) os livros de matemática, por conta do excesso de algarismos, tabelas, gráficos, equações, etc. – elementos descartados, porque não constituem texto em língua natural; b) os livros de português, devido à alta freqüência de textos literários para os quais não obtínhamos os direitos autorais, além da presença de textos poéticos, que são excluídos do CN em virtude da aplicação; e c) os livros de língua estrangeira, pelo motivo óbvio do recorte lingüístico do CN.

Afora esse recorte na representatividade dos campos de conhecimento, esse corpus também foi orientado pela facilidade de acesso ao material. Apesar de contemplar todas as séries escolares (as pastas vão da 3ª série do ensino fundamental ao 3º ano do ensino médio), não existe uma sistematicidade dos domínios. Por exemplo: os textos de geografia constam apenas nas pastas da 6ª e 8ª séries e 2º colegial; já os textos de ciências (ou biologia, para o ensino médio) constam em todas as pastas, e os de historia, na grande maioria do corpus. Portanto, essa distribuição precária de textos por domínio gera a especificidade de que falávamos a respeito do contraste numérico entre o corpus Colégio e as enciclopédias? Apesar de volumoso, o diretório Colégio é mais concentrado em termo de domínio e, por isso, apresenta menos variabilidade léxica.

Por fim, é importante mencionar aqui o fato de que esse corpus Didáticos em geral é o único do CN que seguiu procedimento especial de compilação. Uma vez que os livros didáticos são, em geral, volumosos e, além disso, tendo em vista que a compilação foi empreendida via scanner, estabeleceu-se que os livros seriam recortados em três partes: um trecho do início, um trecho do meio e outro do final de cada obra. Assim, nenhum texto desse corpus é integral, e isso se desvia do perfil geral do CN.

Hoje sabemos que tal procedimento não é condenado pela Lingüística de Corpus, a qual se justifica pelos recortes dizendo possibilitarem representar um texto em sua evolução de dificuldade, i.e., toma-se que à medida que se avança na obra, mais elaborada se apresenta a sua linguagem, aumentando, conseqüentemente o grau de dificuldade da leitura. Enfim, se é uma opção aceitável, a compilação parcial dos textos deve ser generalizada no corpus e, sobretudo, divulgada para o usuário.

Em termos estatísticos, como mostram os gráficos do Anexo II, o corpus Didático apresenta grande concentração dos dados no subdiretório, o qual, a despeito disso, não é significativo quando a

contagem exclui a repetição de palavras. O motivo para isso, como vimos, é a opção pela representação de determinados domínios e o grande volume de textos, por se tratar de livros didáticos.

III.3 Corpus Literário de Cc

Este conjunto de **215** arquivos agrupa:

- seis pastas

E essa disposição dos textos apresenta-se, desde já, como um problema! Não há subdiretórios neste corpus, por exemplo, revelando que a organização deste corpus tem critérios duvidosos, ou nem mesmo os possui. Na verdade, este último caso é o que efetivamente se deu com a compilação dos textos literários do CN.

Como apontamos atrás, a produção literária foi tardiamente inserida no CN. Ela aconteceu num contexto em que se combinava o trabalho de expansão do Corpus e o desenvolvimento do *Módulo de Literatura*[6]. Parte do material disponibilizado no Módulo foi incorporado ao CN, sendo os textos classificados de acordo com as etiquetas recebidas lá. Por isso, temos no Corpus uma pasta para antologias, uma para crítica literária, outra para escola literária e, finalmente, a pasta para resumos. Tais divisões não respeitam, propriamente, uma tipologia de texto, mas sim uma expectativa de consulta, já que o usuário padrão do Módulo – estudantes e vestibulandos – é suposto como aquele que se guia por essas classificações para a sua preparação literária na escola ou no vestibular.

E a prova final de que a organização do corpus não segue um critério objetivo são as duas últimas pastas que ele contém: a pasta “Infantil” e a “Literatura brasileira”. Qual parâmetro utilizado para diferenciar tais textos? E acima de tudo, não é verdade que os textos do Corpus são essencialmente do português do Brasil? Então, uma pasta para literatura brasileira não seria tautologia?

As especificidades desse corpus literário são, no entanto, muito maiores! Vejamos, então, com detalhes, cada uma das categorias:

- Antologia

Nesta pasta, apresentam-se fragmentos de texto, em geral 2 a 5 páginas não-seqüenciais da obra de referência. Não há padrão regular de compilação (ora tem-se a representação de duas ou três obras do autor, ora apenas uma), e isso descaracteriza, por si só, o material considerado antologia. Por fim, dada

[6] O “Módulo de Literatura” é um aplicativo complementar do ReGra, construído pelo Nilc em ___, e que funciona independente da revisão gramatical como um material eletrônico de consulta à literatura brasileira. Para maiores informações, veja-se http://www.nilc.icmc.usp.br/nilc/literatura/bemvindo.htm.

a repetição de obras e autores na pasta Literatura Brasileira, é possível prescindir das antologias no CN.

- Crítica Literária e Escolas Literárias

Sua eficácia no CN é questionável. Os textos dessas duas pastas são esquemáticos, ao estilo das apostilas de cursinho. São pastas pouco representativas, porque contemplam um pequeno número de autores. Por outro lado, têm a vantagem de serem textos com autenticidade de autoria, pois foram elaborados pelo corpo de professores envolvidos na construção do Módulo de Literatura.

- Infantil

Esta pasta tem um grande apelo, pois quer representar o universo infantil e isso é sempre muito difícil no trabalho de construção de corpus escrito. Todavia, o conjunto de textos é inexpressivo (apenas 6 arquivos) e a extensão dos arquivos, muito pequena (em média, 400 palavras, cerca de 1 página de texto). Há de se verificar, além disso, a questão dos direitos autorais dos textos, uma vez que alguns autores representados são contemporâneos.

- Literatura Brasileira

Já mencionamos atrás o fato de ser redundante a nomenclatura desta pasta. A justificativa para tal, contudo, é a de apresentar textos considerados representativos da nossa literatura, e não apenas os tidos como clássicos. Esse ideal, porém, não foi maior que a facilidade de acesso ao texto e, por isso, sem que se avaliasse o grau de representatividade literária, muitas obras foram incluídas na pasta quando da expansão do corpus.
Na pasta, grande parte dos arquivos é referente a fragmentos de textos, em vez de reproduzirem a obra completa. Em muitos casos, inclusive, não há indicação de qual o trecho reproduzido e, neste caso, o seu aproveitamento num corpus rígido quanto a autenticidade, tal como o Lacio-Web, é simplesmente nulo. Em termos de representatividade, esta pasta de Literatura não se mostra um conjunto recomendável: o material não é representativo do autor, do período literário e, por conta da fragmentação sem critérios objetivos, também não é representativo das obras.

- Resumos

Embora não aponte os dados completos de referência, esta pasta tem grandes vantagens. Em primeiro lugar, tal como as críticas literárias, foram textos produzidos pelos organizadores do Módulo. Portanto, a autenticidade dos textos é garantida. Em segundo lugar, esta é uma categoria específica de tipo de texto e é bastante requisitada na composição de um corpus escrito, sobretudo porque há muitos trabalhos que pesquisam a estrutura específica dos resumos. A grande desvantagem que a pasta apresenta, porém, é o baixo número de arquivos (34), que torna o conjunto pouco representativo da literatura.

Enfim, esses são os apontamentos pertinentes na avaliação do material literário do CN. Antes de encerrar este tópico, é preciso ponderar a falsa impressão de ineficácia desse corpus literário. Ressalta-se, pois, o esforço da compilação, que não é tarefa fácil reproduzir as obras da literatura em formato eletrônico, inclusive pelo seu volume. Ademais, ainda que os problemas de representatividade sejam preocupantes, o conjunto dos textos tem permitido a existência da diversidade do CN e, com isso, garantido a base para os testes da ferramenta de revisão, o que significa dizer que ele tem cumprido seu propósito de criação.

III.4 Corpus Técnico e Científico de Cc

Este conjunto de **132** arquivos agrupa:

- duas pastas

O corpus Técnico Científico é bastante pequeno, mas os problemas que os envolve são extremamente superiores! Trata-se de duas pastas – Livro e Livro-Usp (ou Livrusp, como foi nomeado no Corpus) – que não obedecem a critério algum de compilação de textos, exceto o de insinuar-se como textos com alguma cientificidade (ainda que isso se traduza, por vezes, em academicidade).

Na pasta Livro, apenas um arquivo pode ser tomado com um texto técnico-científico, dada a sua aparência de livro extraído de tese; esse texto é o “geologia”. Os outros quatro que compõem a pasta são altamente questionáveis sobre o gênero, sem contar que um desses quatro arquivos repete o texto, mudando apenas o título do arquivo.

Quanto à pasta Livrusp, os problemas começam pela impossibilidade de se recuperar os dados de referência (a autoria e a publicação, dados essenciais, simplesmente não constam dos textos). É

motivo de dúvidas, ainda, o fato de esses textos se inserirem: a) entre o material considerado técnico-científico; b) entre os textos ditos corrigidos; e c) nos títulos do chamado Livro Usp.

Finalmente, o material da pasta encontra-se com muitos problemas de apresentação: há lixo no texto, grandes espaços em branco e trechos com grande concentração de símbolos (equações, por exemplo) ou de escrita estrangeira.

III.5 Corpus Jornalístico de Cc

Este conjunto de **3.379** arquivos agrupa:

- duas pastas gerais: "Esportes" e "Futebol";
- dois subdiretórios: "Revista" e "Jornal";

Já dissemos, aqui, que este é o maior conjunto de textos do Cc, e uma das razões para isso é a presença do jornal *Folha de São Paulo,* com textos de 1994 compilados inteiramente. Mas, à parte essa consideração, o corpus jornalístico merece atenção quanto a outros aspectos.

Iniciando pelas duas pastas gerais, conforme já adiantamos, houve um trabalho especifico do Nilc, em outras épocas, que resultou na compilação particular de textos referentes ao tema esportivo. Um acordo científico ligado ao desenvolvimento do projeto UNL, em 199_, levou os projetistas a introduzirem no Corpus o maior número possível de textos ligados a esporte, especialmente o futebol. Na ocasião, a justificativa da escolha foi a representatividade do tema em se tratando de Brasil, tido como o "país do futebol". Assim, o esforço dessa atividade, guiada pelo critério da aplicação computacional, gerou um corpus sem muitos cuidados na compilação. Dentre os problemas introduzidos com essa perspectiva, o maior deles foi, sem dúvidas, o canal escolhido para a importação dos textos.

Contrariando o critério do registro (escrito em vez de eletrônico), os textos dessas duas pastas foram retirados de *sites* jornalísticos, sem que fosse respeitada a necessidade de anotação completa da fonte. Por conta disso, os textos que lá constam não registram a data de publicação (i.e., a data da divulgação da notícia), a autoria e nem mesmo qual o *site* acessado. Em termos de autenticidade, portanto, essas pastas são absolutamente negligentes.

Quanto à modalidade lingüística, podemos dizer que o desenvolvimento do uso de computadores em rede e o surgimento da Internet introduziram o chamado "registro eletrônico de língua", em oposição aos consagrados registro escrito e registro falado. Mais que uma nomenclatura especial, a idéia de registro eletrônico sustenta que a distinção é pertinente, porque estão envolvidas características especiais de discurso, de estruturação sintática e escolhas lexicais na construção do

texto eletrônico. Dessa forma, uma produção nesse registro, por exemplo, um *site*, difere do registro escrito pela possibilidade do hipertexto, de maneira particular, e pela objetividade, de maneira geral. Portanto, um corpus como o CN, dizendo-se refletir o registro escrito do português, não pode servir-se de textos de *sites*.

Além desses problemas envolvendo a autenticidade e o registro, há dois outros aspectos importantes a se considerar nessas pastas do corpus jornalístico. O primeiro se refere ao produtor dos textos e sua relação com a língua portuguesa; o segundo tem a ver com a formatação dos arquivos.

Um dos canais de pesquisa encontrado para a importação dos textos, especialmente para a pasta de futebol, foram os sites de jornais internacionais (o *New York Times* on-line, por exemplo). Isso trouxe para o Corpus o problema da relação entre o autor do texto e a língua em que escreveu, uma vez que os chamados "correspondentes" nem sempre são falantes nativos do português, o que não garante ao Corpus a autenticidade da língua representada. Mesmo os jornais nacionais eletrônicos costumam manter em sua equipe esses autores-correspondentes os quais, vez ou outra, são estrangeiros com o domínio fluente do português. Ocorre que, pelo critério da língua representada no corpus, o autor dos textos deve ser falante nativo, porquanto sua produção pode servir de base para pesquisas sobre a estrutura sintática da língua, por exemplo. No corpus jornalístico, então, os textos assinados por esse tipo de indivíduos têm de ser excluídos, a fim de que se garanta ao usuário que o CN é uma base de textos produzidos em língua portuguesa por falantes do português do Brasil.

O segundo aspecto que devemos comentar diz respeito ao padrão de formatação dos arquivos. Na verdade, o problema que vamos relatar envolve o corpus jornalístico como um todo e não só as pastas gerais. Trata-se, em primeiro lugar, da presença de mais de um texto em cada um dos arquivos e, em segundo, da falta de padronização dos dados de referência em cada um dos arquivos.

No primeiro caso, o Corpus apresenta problema, porque o acúmulo de textos nos arquivos não permite a contabilização correta dos textos, por exemplo, falseando estatísticas a seu respeito. O ideal seria que um arquivo contivesse apenas um texto, ainda que este fosse curto. No segundo caso, o problema que emerge é a impossibilidade de manipular automaticamente os dados de referência dos textos, uma vez que houve falta de padrão na anotação dessas informações quando da compilação. Embora pareça trivial para os leigos, o simples fato de dispor *ordenadamente* a manchete da notícia, a autoria e a data da publicação, por exemplo, torna isso um procedimento especial para o processamento automático dos textos, já que são possíveis as buscas por padrões. Um efeito prejudicial dessa negligência da compilação é o esforço que os envolvidos no Lacio-Web terão para aproveitar os textos jornalísticos do CN, já que terão de percorrer cada página de cada arquivo, a fim de recuperarem, manualmente, os textos lá existentes.

Passemos, agora, à análise dos demais materiais desse corpus jornalístico. No que se refere ao subdiretório Revistas, a principal consideração a fazer é sobre a sua pouca representatividade. Quer o número de textos, quer a época de publicação ou mesmo a fonte institucional são inexpressivos. O corpus Revistas conta com quatro pastas, cada qual reservada a uma fonte institucional específica: a revista *Época*, a *Istoé*, a *Globo Rural* e a *Veja*. São periódicos do jornalismo ordinário, mas regional, i.e., produzidos na região sudeste do Brasil, com toda a sorte de particularidade discursiva que isso produz. Além disso, cada uma dessas pastas, embora represente publicações diárias e, por isso, um conjunto bastante extenso de textos, possui menos de 5 arquivos, o que resulta dizer que não refletem a produtividade informativa das revistas compiladas (apenas a revista *Época* tem número superior de arquivos no corpus: 9 no total). Finalmente, não houve a preocupação de cobrir um período suficientemente amplo de publicação, o que permitiria, por exemplo, o desenvolvimento de pesquisas lingüísticas de cunho histórico.

Por essas razões, esse subdiretório merece um esforço concentrado no sentido de alimentar as pastas com um número maior de arquivos, ainda que se restrinjam as fontes institucionais. Há de se ressaltar, além disso, que as considerações sobre a formatação dos arquivos mencionada atrás também valem para esse subdiretório, o que implica um trabalho de subdivisão de muitos de seus arquivos que acumulam textos.

Encerrando esse corpus jornalístico, aparece o maior conjunto de textos: o corpus dedicado aos jornais. E a primeira característica que chama a atenção em relação a ele é a sua extensão, que ultrapassa, sobremaneira, todos os demais conjuntos do corpus corrigido do CN. O subdiretório Jornal é composto por 4 pastas, referentes aos seguintes periódicos: *Jornal do Brasil* (pasta Brasil), *Folha de São Paulo* (pasta Folha), *The New York Times* (pasta de mesmo nome) e *Jornal da Usp* (pasta Usp).

De pronto, essa composição nos revela que o subdiretório sofreu a influência de uma construção de corpus sem o aparato do estudo lingüístico. Por isso, uma das conseqüências foi a apresentação conjunta de periódicos de cunho puramente informativo e de grande circulação (ex.: Folha de São Paulo - FSP) com aqueles de caráter mais científico e de circulação restrita (ex.: Jornal da Usp). Tal disposição é considerada inadequada, uma vez que neutraliza a característica essencial do gênero textual (por exemplo, a oposição informatividade vs. cientificidade), privilegiando, com isso, apenas o caráter geral dos textos, que é, no caso, a "divulgação". Para os compiladores do CN, então, qualquer texto de divulgação seria incluído no corpus jornalístico, sem distinção quanto ao seu caráter acadêmico-científico, por exemplo.

Ainda com respeito à constituição do corpus, é importante comentar a presença inadvertida do jornal *The New York Times*. Considerando que o CN se dedica à compilação de textos do português,

manter esse jornal americano no Corpus não tem qualquer fundamento! Além disso, a pasta dedicada aos textos desse jornal tem 9 arquivos, armazenando parcialmente apenas um mês de publicação, o que significa dizer que o jornal está muito pouco representado no Corpus.

A questão da representatividade do periódico no CN, aliás, é o principal problema da pasta Usp. São apenas 2 arquivos existentes, o que frustra qualquer expectativa de pesquisa que pretenda valer-se desse tipo de texto. A pasta Usp, além disso, devido a problemas de compilação, tem falhas enormes na indicação dos dados de referência. A indicação sobre a autoria dos textos, por exemplo, é nenhuma. Seria preciso recuperar a publicação daquela época (1994) e buscar, manualmente, pelas matérias compiladas, a fim de que essas referências fossem resgatadas. A pouca representatividade da pasta, no entanto, inviabiliza todo esse esforço e sugere, em contrapartida, que nova compilação seja promovida.

Aliás, num trabalho de construção da base jornalístico-científico, outro aspecto tem de ser corrigido: a variedade das fontes. Da forma como aparece no CN, a pasta Usp está completamente isolada, pois não tem um conjunto textual de mesma natureza, cuja especificidade seja, apenas, a instituição (por exemplo, jornal da Usp vs. jornal da Fapesp). Se se mantiver apenas o jornal da Usp, é necessário que haja, ao menos, a representatividade do mesmo, de modo que o estilo jornalístico-científico possa ser evidenciado num trabalho de pesquisa.

Esse contraponto da fonte, por sua vez, é oferecido pela pasta Brasil em relação à Folha. Apesar de o conjunto ser inferior em número de textos, a pasta Brasil é representativa do periódico, somando 481 arquivos no CN.

O *Jornal do Brasil* que aparece no Corpus são textos publicados em 1996, especialmente no primeiro semestre. E tal como ocorre com a FSP, os textos foram organizados em diretórios individuais, referentes a cada uma das seções do jornal (os chamados "cadernos"), oferecendo um padrão muito preciso de organização textual. No que diz respeito aos problemas encontrados nessa pasta, apenas se destaca a ausência da autoria de algumas matérias e o acúmulo de textos em cada um dos arquivos.

Finalmente, a pasta FSP, como já dissemos, aparece no Corpus com o maior número de textos: são 22 diretórios referentes aos cadernos do jornal, num total de 3.360 arquivos. Um único diretório – o caderno "Folha Ciência" – contraria o padrão dos arquivos desse corpus, uma vez que reúne apenas 3 arquivos, sendo que um deles tem somente uma página de texto. Nos demais diretórios, há uma extensão considerável dos arquivos, fruto da inclusão de mais de um texto (i.e., matéria jornalística) em cada arquivo. Outra exceção ao padrão oferecida pela Folha Ciência é a data da publicação, com textos de 1996, quando os demais datam de 1994.

A extensão do FSP no Corpus gera um sério problema de balanceamento, conforme os dados do Anexo II podem demonstrar. A distância dessa pasta em relação a qualquer outra do corpus corrigido é evidente e deve ser considerada num trabalho de análise estatística, conforme apontaremos na seção IV deste documento.

III.6 Corpus Universitário de Csc[7]

Este conjunto de **184** arquivos agrupa:

- sete subdiretórios

Esses conjuntos são os seguintes: Artigo, Aula, Dissertação, Projeto, Qualificação, Relatório e Tese. Na nossa avaliação, excluímos o subdiretório Aula, pois ele se compõe de anotações não publicadas e, por isso, não relevantes para os propósitos do Lacio-Web.

Conforme é possível pressupor, o corpus universitário é constituído por textos relacionados à atividade acadêmica. A separação dos subdiretórios, portanto, segue uma tipologia de texto baseada nos diferentes documentos produzidos no meio científico. No interior de cada subdiretório, a classificação dos textos se dá pela área de conhecimento, as quais, por sua vez, nomeiam cada uma das pastas (ex.: subdiretório "Dissertação" – pastas "Biológicas", "Exatas" e "Humanas").

Contudo, essa separação dos subdiretórios não é sistemática: há casos em que apenas uma área de conhecimento tem textos no corpus. Isso nos leva, mais uma vez, a considerar que, embora tenha havido preocupação em organizar o corpus segundo uma tipologia (de campo de conhecimento, no caso), o critério efetivamente seguido foi o da facilidade de aquisição do material. Um exemplo típico dessa orientação é o subdiretório "Artigo", que não separa os textos por campo de conhecimento, mas, numa pesquisa aos textos propriamente ditos, reconhecemos tratar-se de material produzido no âmbito da Física. No total, são 11 artigos voltados para a mesma área.

A avaliação desse corpus, então, precisou alcançar esse nível dos textos em si. E, para dar conta dos resultados apurados, torna-se mais adequado tratar os subdiretórios separadamente.

- Artigo

Além do fato relatado acima envolvendo a concentração dos textos a um único campo do conhecimento, outro aspecto relevante desse subdiretório é a absoluta falta de referência dos textos. De maneira sistemática, os artigos não apresentam indicação sobre a autoria, tampouco da publicação.

Salvo uma consulta ao compilador, será impossível recuperar esses dados, o que é essencial para a documentação.

- Dissertação

Esse subdiretório é divido nas três pastas relativas aos campos de conhecimento: Exatas, Humanas e Biológicas. De imediato, a falta de equilíbrio, quanto ao volume, entre as três pastas é evidente, sendo a Exatas a pasta com o maior número de textos (29), seguida da Humanas (3) e da Biológicas (2).
As pastas Biológicas e Humanas, além de inexpressivas, contêm textos sem indicação dos dados de referência, dentre os quais, a autoria. Mas, o importante a ser mencionado, aqui, é a inclusão de textos parciais, i.e., não se trata da dissertação integral, mas de capítulos ou de seções do trabalho (ex.: h01, que é o 5º capítulo de uma dissertação na área de humanas).
Quanto à pasta Exatas, pode-se dizer que tem representatividade, dado o volume de textos compilados. No entanto, a representatividade da pasta se restringe a esse aspecto, pois os textos lá contidos são dedicados à área da Matemática, da Computação e da Física (portanto, limitação de campo específico de conhecimento) e, em sua totalidade, foram produções oriundas de uma única instituição universitária – a USP.
Não parece acontecer no Exatas a compilação parcial dos textos, tal como ocorreu com as duas outras pastas. E, em se tratando de referências sobre os trabalhos, os textos estão marcados quase completamente, faltando apenas a indicação da autoria. Contudo, o problema particular dessa pasta é a falta de padrão da formatação dos textos, que demonstram heterogeneidade na apresentação dos cabeçalhos (uns são ostensivos, outros omitem dados importantes dos trabalhos, tais como o nome do orientador, a unidade em que ocorreu a defesa, etc.). Além disso, como é próprio do campo, há os espaços referentes à presença de tabelas, figuras, equações, fórmulas, etc. A padronização desses elementos, porém, foi negligenciada quando da compilação. Ou seja, uma vez que não são considerados textos, esses elementos são retirados do material, mas, em vez de simplesmente desaparecerem do texto, é fixada uma marca do tipo de informação ocultada a fim de que o leitor-pesquisador possa obter a linearidade discursiva do texto. Essa marcação, justamente, é heterogênea nos arquivos da pasta, o que prejudica, por exemplo, o aproveitamento dos textos num corpus anotado.

- Projeto

[7] Conforme dissemos na Introdução, dado o escopo deste documento, os demais conjuntos do corpus semicorrigido, e o corpus não-corrigido em sua totalidade, não são considerados.

O subdiretório contém a pasta Exatas e a Humanas. Ocorre que, da mesma forma que aconteceu em Dissertação, a pasta Exatas é bastante superior em volume: 17 arquivos contra apenas 1 da Humanas.
O problema mais evidente desse subdiretório é a total ausência dos dados de referência. Soma-se a isso, o fato de que, além da pouca representatividade dos campos de conhecimento, os textos desse corpus são quase inteiramente voltados para a área da Física, representando muito pouco, então, o próprio campo das Exatas.

- Qualificação

Esse subdiretório parece ser o mais problemático! Compõe-se de apenas uma pasta – Exatas – e esta, de 5 arquivos somente.
O grande problema, contudo, é o fato de não serem textos integrais e, dentre os 5 arquivos, alguns serem textos repetidos. Os textos compilados, em seu cabeçalho se auto-intitulam “mini-dissertação”.

- Relatório

O subdiretório separa pastas para Exatas e Biológicas, mas, como em outros casos, a proporção dos arquivos é de 50 textos contra 1, demonstrando a total falta de equilíbrio do corpus.
Nos relatórios da Exatas, além da já tradicional ausência de dados de referência, o que chama a atenção é a presença de textos mais burocráticos do que propriamente técnicos. Num corpus de tipologia mais refinada de textos, seria o caso de separar esses textos: de um lado, os relatórios técnico-acadêmicos e, de outro, os técnico-administrativos.
Finalmente, outro aspecto que se destaca nesse corpus é a existência de textos em inglês, o que deve ser completamente descartado num novo corpus.

- Tese

Os textos desse corpus estão reunidos numa única pasta – Exatas – e somam 56 arquivos. Apesar desse volume significativo, os arquivos não são expressivos uma vez que não foram compilados de forma integral. Em geral, o conteúdo do corpus é o último capítulo das teses, justificando o número reduzido de páginas dos arquivos (em média 10 páginas).

Há, ainda, o problema da concentração dos textos numa única área (em geral, da Física e Computação). Somado a isso, é evidente a falta de padronização do formato dos arquivos. Em vista disso, o corpus Tese não deve ser alvo de aproveitamento no Lacio-Web e, indo mais além, não se presta a uma pesquisa lingüística sofisticada.

Conforme apontamos no início deste documento, o corpus semicorrigido se compõe de textos no intervalo entre a escrita espontânea e a escrita formal revisada. Por conta disso, é natural a falta de padrão na apresentação dos textos. Assim, a definição de uma formatação padrão deve ser preocupação dos responsáveis pela compilação do corpus, que precisarão assumir um formato de acordo com as expectativas de utilização do corpus (o desenvolvimento de ferramentas e de periféricos do processamento automático, o desenvolvimento de etiquetas para anotação de corpus, etc.).

No que diz respeito aos dados de referência, tantas vezes considerados aqui como uma falha da compilação, é preciso dizer que foi um procedimento de trabalho adotado na época. Quando os corpora semi e não-corrigido foram desenvolvidos, assumiu-se que a autoria dos textos não seria um dado relevante. Além disso, a fim de alcançar a autorização de uso dos trabalhos acadêmicos, uma das condições oferecidas aos fornecedores foi, justamente, a ocultação no Corpus da autoria dos seus textos. Na prática, outros dados dos trabalhos passaram a não ser revelados, o que, hoje, se demonstrou ser imprescindível na composição de corpus.

III.7 Sumário

A construção do *Corpus-Nilc* (CN) foi um empreendimento decorrente da ferramenta de revisão desenvolvida pelo NILC em parceria com a ITAUTEC/PHILCO a partir de 1993 – o revisor *ReGra*. Inicialmente projetado para suportar exigências específicas do trabalho de revisão, o CN assumiu duas tarefas fundamentais: de um lado, servir de base para os testes de desempenho do ReGra e, de outro, fornecer parâmetros para os estudos lingüísticos requisitados pelo Revisor. Assim, a construção do CN pautou-se na afinidade necessária com a aplicação computacional a que se vinculava, em vez de basear-se numa orientação teórica adequada, contemplando a escolha objetiva de critérios de construção de corpus e de seleção e organização dos textos.

Em termos gerais, o Corpus foi alimentado de acordo com dois parâmetros de seleção dos textos: a facilidade de aquisição do material e a pertinência do mesmo para o Revisor. Desse modo, construiu-se uma ampla base textual que, em maio de 2000, contava com 35.215.783 ocorrências.

No decorrer dos anos de pesquisa, o CN sempre mereceu alguma atenção específica, especialmente quanto à inclusão de novos textos e à reformulação de sua estrutura de organização interna. Esses trabalhos resultaram na apresentação de um corpus divido em 3 diretórios macros, em que se distingue o grau de adequação da escrita ao padrão gramatical do português culto: o corpus corrigido, no topo da lista, reservado aos textos considerados livres de erros, porque tomados como tendo sido revisados (ex.: os livros didáticos); o corpus semicorrigido, num nível intermediário desse comportamento lingüístico (ex.: teses); e o corpus não-corrigido, em que se reúnem textos de escrita espontânea ou que não tenham sido alvo de revisão lingüística consistente (ex.: redações de vestibulandos).

Além dessa divisão macro, o CN ainda apresenta separação dos textos em diferentes níveis de classificação, mas, desta vez, não há uma organização criteriosa, nem mesmo segundo as exigências da aplicação, como foi o caso da divisão apresentada acima. Os textos são alocados em pastas que se distinguem entre si pelo tipo de texto, pelo tipo do assunto ou pela área de conhecimento. E, porque não houve um estudo precedendo essa organização, alguns problemas surgem, como é o caso da existência de pastas isoladas (i.e., fora de um subdiretório específico na classificação interna dos diretórios macros) ou da concentração de textos de pastas gerais, quando o desejável seria a separação dos mesmos em pastas mais específicas (de assunto ou de tipo de texto, por exemplo).

Observações pontuais como essas, com o passar do tempo, tornaram-se alvo de constantes incômodos, mesmo para os pesquisadores do NILC envolvidos com estatísticas de corpus e recuperação de informação. Disso decorreu a expectativa de se construir um novo corpus a partir desse CN existente. Concretizado pelo projeto LACIO-WEB, que visa ao desenvolvimento de um corpus do português do Brasil para estar disponível na Internet, a proposta de reformulação do CN passou a exigir uma ampla avaliação, a partir da qual as diretrizes de aproveitamento do material existente poderiam ser efetivamente elaboradas. Eis, pois, o objetivo deste documento.

Na avaliação em questão, diversos pontos foram levantados dando conta do estado atual do CN, no geral, e de maneira específica, do corpus corrigido (em sua totalidade) e do corpus semicorrigido (uma única pasta). Esse recorte, por sua vez, deveu-se ao fato de que esses são os diretórios de aproveitamento potencial do CN para o Lacio-Web.

No relato das condições em que se encontra o CN, cada um dos níveis de alocação dos textos foi examinado, possibilitando que, numa pesquisa com base em algum dos corpora, os analistas pudessem conhecer as limitações desse material. Essas restrições, conforme argumentamos no decorrer desta seção, são as seguintes:

♦ classificação e catalogação de textos: a classificação dos textos é problemática, pois o Corpus foi construído sob demanda. Verifica-se a necessidade de uma organização textual de acordo com uma tipologia (uma reivindicação, aliás, presente em documento antigo do próprio NILC). Dentre os problemas vinculados à falha na organização e catalogação dos textos, estão os seguintes casos: a) os códigos, do corpus jurídico; b) as enciclopédias, do corpus didático; c) o corpus literário; d) os textos sobre esportes, do corpus jornalístico.

♦ representatividade: algumas pastas são muito pouco representativas, i.e., não contam com exemplares quantitativamente suficientes do título que as reúne num diretório. Isso foi observado com mais ênfase no corpus técnico-científico e no corpus universitário. Em certa medida isso também se aplica ao corpus literário, já que é possível considerar falha a representatividade do corpus segundo o autor, ao conjunto das obras, ao período literário, etc. Um caso extremo da pouca representatividade desse corpus é o diretório "Infantil", inexpressivo em termos de número de amostras.

Por outro lado, pode-se falar que é representativo no CN o corpus didático, pela variedade de textos segundo as séries escolares e aos campos de conhecimento. No entanto, esse corpus apresenta o problema da amostragem (veja mais, abaixo), o que torna relativo esse atributo de representatividade. A respeito do corpus jornalístico, quantitativamente o mais representativo do CN (seguido do corpus de redações – corpus não-corrigido), essa prerrogativa deve ser atenuada pelo fato de que há uma grande concentração das fontes dos textos, em que o jornal *Folha de São Paulo* e o *Jornal do Brasil*, ambos com 1 ano de publicação compilado, são as únicas fontes significativas. Falta representar as demais categorias previstas no corpus, por exemplo, as revistas.

♦ compilação: alguns tipos de textos tiveram compilação irregular em relação ao padrão de amostragem aplicado em quase todo o CN. Em geral, o Corpus contém textos integrais e seqüenciais das obras. Isso não é o caso, porém, em alguns itens do catálogo do Corpus. O corpus didático apresenta seus textos recortados por trechos: uma parte do início, uma do meio e outra do fim de cada obra. Por sua vez, no corpus literário (em especial o diretório "Literatura Brasileira") os trechos compilados não são seqüenciais da obra, agravando o problema da amostragem integral dos exemplares do Corpus. Finalmente, em alguns casos o corpus universitário também se inclui nesse paradigma de problemas, uma vez que em vez de teses integrais, por exemplo, o que se tem no CN são alguns trechos do trabalho, em geral o último capítulo. Há, pois, uma necessidade de adequação desses casos ao padrão de amostragem que se procura para o CN, ou, simplesmente, a divulgação dessa característica especial para o usuário dos corpora.

♦ particularidades da aplicação: compilação sob demanda dos textos ocasionou a existência de pastas "soltas" no Corpus. É o caso da pasta "Futebol", do corpus jornalístico. Esse material se desvia

do padrão de catalogação do CN, que privilegia a reunião dos textos por ordem de domínio e de gênero/tipo textual em detrimento de pastas especializadas segundo o assunto. Esse problema acena, então, para a necessidade de inserção criteriosa de novas pastas, diretórios e subdiretórios.

♦ acúmulo de textos: outra discrepância encontrada entre os repositórios dos textos do CN foi o acúmulo de textos em um único arquivo. Embora a reprodução integral dos textos fosse um princípio adotado na construção do Corpus, em alguns casos tal procedimento se mostrou insatisfatório, uma vez que oculta informações importantes dos textos e, por vezes, até mesmo os próprios exemplares textuais. Nessa categoria estão, por exemplo, os textos das enciclopédias e os jornalísticos, que agrupam diversos pequenos textos num único arquivo. Para o caso dos jornais, esse problema é atenuado pelo fato de que os arquivos são separados por data de publicação, incluindo-se em cada um deles todos os textos, de todos os cadernos veiculados naquela data. Para as enciclopédias, no entanto, não há sequer esse critério de separação dos arquivos, que foram divididos à medida que atingiam grande volume de dados. Nota-se, então, que numa catalogação mais refinada desses textos será preciso individualizar as amostras, uma vez que elas são variáveis em assunto e em campos do conhecimento, especialmente. Esse trabalho contribuirá para uma organização e classificação textual mais refinada do corpus, permitindo a busca por textos específicos os quais, hoje, encontram-se embutidos em categorias muito gerais de tipologia textual.

♦ nomeação dos arquivos: um cuidado especial também deve ser dedicado à nomeação de cada amostra textual de um corpus. Na nossa pesquisa, detectamos um número significativo de repetição de textos, i.e., textos iguais com rótulos diferentes. Isso se dá, especialmente, no corpus literário (vejam-se textos idênticos entre os subdiretórios "Antologia" e "Literatura Brasileira") e no corpus técnico-científico. Tal deficiência, se não identificada e solucionada, gera uma falsa impressão quantitativa do corpus, além de viciar a catalogação textual.

♦ direitos autorais: ainda que o perfil do CN tenha lhe permitido contar com textos sem a concessão dos direitos autorais, para um projeto tal com o Lacio-Web, que disponibilizará o corpus para acesso gratuito e irrestrito, é imprescindível a autorização de uso de cada material compilado. No CN, a pouca compreensão desse fator à época de sua construção ocasionou diversas falhas de anotação dos dados de referência dos textos. Muito do material que representa o corpus universitário teve ocultada a autoria, por exemplo; alguns textos do corpus literário não contam com informações de publicação; e o corpus jurídico não informa quais as suas fontes. Todas essas lacunas demonstram que o CN, de fato, não foi projetado para servir como referência de pesquisas fora do NILC. Numa proposta diferente de trabalho, contudo, esses campos deficientes devem ser reformulados, provendo-se de extensa anotação dos dados bibliográficos dos textos, dos dados lingüísticos internos, além de

um Termo formal de Autorização de Uso para os textos que não forem, legalmente, de domínio público.

Enfim, esses são os aspectos mais expressivos detectados na nossa avaliação do CN. Eles dão conta do **estado atual do Corpus**, em setembro de 2002.

Até essa data, o CN conta com **35.197.539** ocorrências. Nem todas elas, porém, representam palavras válidas para o português, ao contrário do que anunciam outros relatórios sobre o Corpus. Essa é, pois, a questão discutida a seguir.

Além desse tema, a seção IV também apresenta a distribuição desse número de ocorrências pelos corpora do CN, tocando, então, no problema do balanceamento do Corpus.

No que concerne ao escopo deste documento, o conjunto de textos analisados somam 34.092.630 ocorrências, e estão distribuídas conforme o quadro a seguir. Dele foi excluído o conjunto de textos jornalísticos, já que sendo muito maior que qualquer outro (26.623.406 de ocorrências), ele não permitiria a visualização das demais categorias de textos.

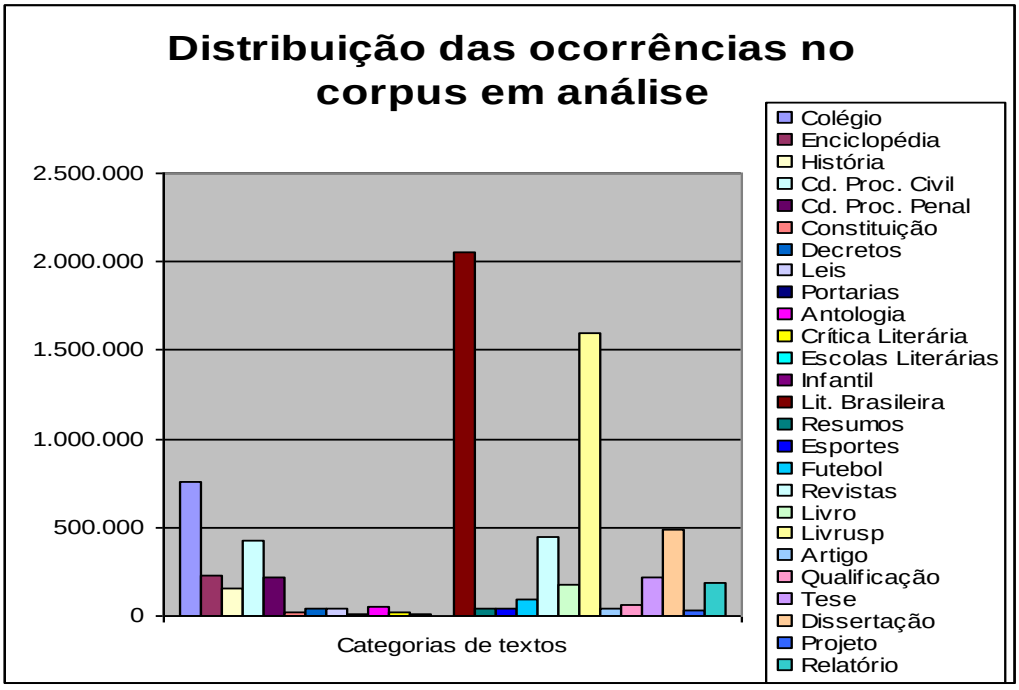


Fig. 1 Quadro de distribuição de ocorrências

IV. Contagem de Freqüência e Distribuição da contagem pelos corpora do CN

De tempos em tempos o Nilc promove uma contagem de freqüência no CN e é ela que substancia os números de ocorrências divulgados sobre o Corpus. Essa contagem, no entanto, é passível de grandes armadilhas! E porque as lacunas do instrumento influenciam nas decisões de análise, convém reportarmos aqui alguns fatos importantes e outros curiosos para os quais o usuário do contador de freqüências do CN deve estar atento.

Anunciamos atrás que o CN apresenta, em 2002, **35.197.539** de ocorrências. Mas esse número precisa ser descrito com mais clareza, uma vez que, nesse volume, está sendo incluído, além das palavras, também as não-palavras do português.

Antes de qualquer coisa, é importante mencionar que esse montante de aproximadamente 35 milhões de ocorrências contém grande volume de repetições. Se subtrairmos as ocorrências repetidas, o Corpus passa a somar **340.016** ocorrências, reduzindo em **99,04** % o volume determinado para o tamanho do Corpus em palavras. Ou seja, apenas **0,6** % do CN é, efetivamente, o número de ocorrências diferentes no Corpus. Em outras palavras, ainda, podemos dizer que as ocorrências não-repetidas do Corpus representam somente **0,96%** dos 35 milhões detectados na contagem geral.

O Apêndice III apresenta um recorte dessa contagem de freqüências produzido em setembro de 2002[8]. Para efeito de análise, retiramos um extrato desse Apêndice, que compõem as Tabelas estudadas nesta seção. Por essa amostragem, poderemos detectar, de um lado, algumas limitações do levantamento estatístico e suas conseqüências para a determinação do número de palavras do CN, e de outro, alguns casos que demonstram os reflexos do estado atual do Corpus para uma análise lingüística.

No que diz respeito ao levantamento promovido, cumpre salientar a presença de um número significativo de itens que se constituem em repetições de letras sem qualquer unidade lingüística do português. Vejamos a tabela abaixo:

Posição	Ocorrência	Freqüência
66.438	foe	13
66.430	fjortoft	13
150.861	àa	2
150.862	aab	2
334.291	uuêi	1
334.292	uuguai	1
334.293	uum	1
334.299	üüüüüüüüüüüüüüüü	1
334.300	uvaia	1

[8] A totalidade desse material está disponível no site do Nilc, bem como o extrato parcial que compõe o Apêndice III.

Tab.1: Ocorrências de não-palavras no Corpus

Elementos como os que aparecem na Tabela 1 têm uma caracterização para a qual o contador de freqüência utilizado não está capacitado para eliminar, uma vez que não temos um padrão seguro das combinações silábicas aceitáveis no português do Brasil (as siglas e abreviaturas, por exemplo, são construções em que as letras se repetem, cf. AABB – Associação de Amigos do Banco do Brasil). E, ainda que nos fosse possível utilizar o léxico do ReGra como filtro nessa operação, tal procedimento nos impediria de aceitar as ocorrências válidas surgidas no Corpus, como é o caso de "AABB". Assim, na versão final da contagem de palavras/freqüência, o programa conseguiu apenas informar o número de ocorrências não-repetidas do Corpus, desconsiderando o fato de serem ou não palavras do português do Brasil.

A Tabela 1 também nos permite verificar que, de fato, há diversos itens não-lingüísticos na relação de contagem. Isso nos faz concluir que o CN tem 35 milhões de **ocorrências** e não de palavras[9]. Ademais, a análise de outros dados (cf. tabelas 2 e 3) também nos possibilita dizer que essas ocorrências, em muitos casos lingüísticas, não são válidas para o português. Nesse caso, além de aquele ser um número que se relaciona a ocorrências, ele não pode ser atribuído unicamente ao português.

No CN, sobretudo pela presença do corpus não-corrigido, são muito comuns os erros de ortografia e as falhas de digitação. Esse fato também sofre as conseqüências do trabalho não-revisado de compilação dos textos, especialmente quando a atividade utiliza o scanner, que é um excelente instrumento de inserção de lixo nos textos em formato eletrônico.

No trecho final da relação de palavras/freqüência, ou seja, as ocorrências de mais baixa freqüência no Corpus, os itens que mais aparecem são os que exemplificam as faltas na digitação, na compilação, e na ortografia das palavras. Tal é o que demonstra a Tabela 2:

Posição	Ocorrência	Freqüência
67.096	niveis	13
150.864	aadequação	2
150.873	aaroun-el-raschid	2
155.960	basea-se	2
155.974	básico-fashion	2
167.626	espritualista	2
167.656	esqui-mó	2
268.166	infidelity	1
334.305	uversitária	1

[9] Esta distinção, ocorrência vs. palavra, não equivale à diferença entre "tipo", "forma" e "token" empregados, mais comumente, em corpus lematizados. Aqui, se uma construção "ocorreu" no Corpus, então ela é considerada "ocorrência". Não quer dizer, contudo, que essa ocorrência se constitua uma palavra, um tipo, uma forma ou um token – classificações que mereceriam um grau sofisticado de análise.

Tab. 2: Ocorrências de baixa freqüência no CN

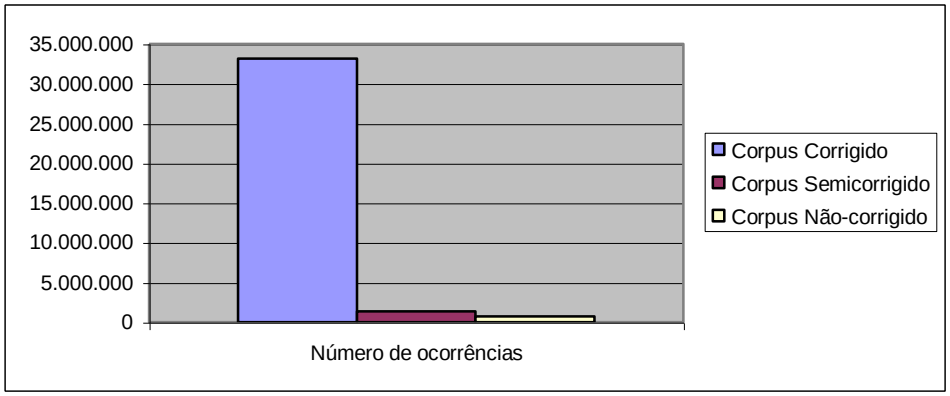
Podemos observar a falta de acentuação, a ausência de letras e pontuação, a inversão de caracteres, etc. Acrescenta-se a esses itens, os casos de neologismo, de erro na grafia de termos técnicos e de estrangeirismos.

Passemos agora, para o segundo aspecto que podemos discutir a partir desse extrato de análise. Para tanto, vejamos os dados sobre os itens de mais alta freqüência no Corpus, compreendidos na tabela 3 abaixo:

Posição	Palavra	Freqüência
1	de	1.701.990
2	a	1.243.051
29	página	100.187
31	editoria	97.283
42	brasil	61.631
47	paulo	57.990
55	governo	44.405
64	folha	42.152
76	us	37.138

Tab. 3: Ocorrências de alta freqüência no CN

Em conjunto com essas informações, vejamos também a distribuição das 35.197.539 ocorrências do Corpus pelos corpora corrigido, semicorrigido e não-corrigido:



Graf. 1: Distribuição das ocorrências do CN

Traçando um paralelo entre a Tabela 3 e esse gráfico acima, é fácil perceber porque determinados fatos lingüísticos foram gerados. Nota-se, por exemplo, que as palavras “folha”, “paulo” e “brasil” estão entre as 64 ocorrências de maior freqüência. Ora, isso é natural, uma vez que o corpus jornalístico é preponderante no CN, e lá o nome dos jornais *Folha de São Paulo* e *Jornal do Brasil* se repete em aproximadamente 5.000 arquivos!

O CN não é um corpus balanceado, seja por extensão da amostragem, seja por assunto, por tipo ou gênero de texto, etc. Dessa forma, o resultado da contagem de freqüência torna-se viciado pelo desequilíbrio dos textos e faz surgir indicações lingüísticas muito mais fiéis ao perfil do Corpus do que comportamento lingüístico do "escrevente" do português[10].

É curioso observar, por exemplo, que a palavra "governo" ocupa o 55º lugar na lista de freqüência, o que pode ser explicado pelo caráter dos textos jornalísticos do CN, muito mais tendencioso a análises políticas do que policiais, por exemplo. Soma-se a isso o fato de que, no Corpus, há os textos oficiais legislativos, o que aumenta a freqüência de "governo".

Já o termo "us", em 76ª posição na tabela, se refere ao símbolo da moeda americana. Trata-se de um dado interessante da listagem de freqüência, uma vez que indica o quão a palavra é citada nas publicações brasileiras, sobretudo jornalística.

E esses casos se observam em todo o CN. A palavra "droga", por exemplo, está na 2.249ª colocação da contagem de freqüência (ocorrendo 105.835 vezes no Corpus). Esse fato não seria relevante se não soubéssemos que um número significativo de redações, do corpus não-corrigido, versasse exatamente sobre esse tema. A alta freqüência de "droga" no CN deve-se, pois, ao privilégio que a palavra teve na composição de textos do Corpus.

A questão do balanceamento de corpus é fundamental para uma análise de dados. E sendo deficiente nesse aspecto, hoje o CN tem se demonstrado, enfim, apenas um banco de texto para referência na construção de corpora mais específicos. A título de ilustração, podemos refletir sobre uma prática que tem se tornado comum no grupo.

Recentemente, o Nilc produziu 3 corpora distintos, cada qual com a finalidade de dar suporte às ferramentas computacionais em desenvolvimento. Tem-se, então, o PROBI, corpus elaborado para testar o desempenho do ReGra; o corpus jornalístico, recortado com o fim de treinar etiquetadores; e o corpos científico (C-C), extraído para análise da estrutura esquemática desse tipo de discurso[11].

A existência desses corpora específicos demonstra que o CN armazena textos relevantes, já que podem ser aproveitados por uma diversidade de pesquisas em PLN. Por outro lado, no instante em que nos deparamos com o PROBI notamos que a função inicialmente projetada para o CN não tem mais

[10] "Escrevente" em oposição a "falante", numa referência ao usuário-alvo do CN.

[11] Para mais informações, cf. Martins, R.t. (2002) *PROBI: um corpus de teste para o revisor gramatical ReGra.* NILC-TR-02-10, 7p.; Aires, R.V.X.; Aluísio, S.M. *Criação de um corpus com 1.000.000 de palavras etiquetado morfossintaticamente*. Série de Relatórios do NILC. NILC-TR-01-8, Outubro 2001, 14p.; e *Feltrim, V.D.; Nunes, M.G.V.; Aluísio, S.M. Um corpus de textos científicos em Português para a análise da Estrutura Esquemática. Série de Relatórios do NILC. NILC-TR-01-4, Julho,2001*, todos disponíveis em http://www.nilc.icmc.usp.br/nilc/publications.htm#TechnicalReports.

efeito, uma vez que o ReGra agora tem outro corpus para testes. Por sua vez, a existência do C-C prova que o balanceamento do CN é, de fato, uma ausência lastimável!

Há de se mencionar, finalmente, que a desambiguação dos itens do Léxico, utilizada no processamento sintático do ReGra, confia na relação de freqüência que hora abordamos. É essencial, portanto, que as falhas de balanceamento e de compilação estejam previstas na programação do sistema, a fim de que os problemas descritos aqui não se propaguem para o funcionamento do Revisor.

V. Perspectivas

As discussões desenvolvidas até aqui, acompanhadas dos apontamentos envolvendo o estado atual do CN, nos levam a assumir que, atuando na reformulação e aproveitamento de conteúdo do Corpus, o *Projeto Lacio-Web* (LW) deve ter em conta os seguintes aspectos:

- visando a atender as exigências da documentação de material imposta ao LW, será preciso empenhar-se na obtenção da autorização de uso junto aos editores, autores, instituições diversas, os quais detém os direitos autorais dos textos que nos interessa. Até esta data (01/2003), os acordos nesse sentido já nos permitem aproveitar o jornal *Folha de São Paulo*, material existente no CN. Outros textos podem se juntar ao jornal, tais como:

Jornal do Brasil	Livros didáticos
Enciclopédias	Literatura

- textos de domínio público são de aproveitamento imediato pelo LW. Estão nessa categoria alguns conjuntos do CN, tais como:

Leis, Portarias, Ofícios, Decretos, etc.	Literatura: textos publicados há mais de 15 anos
Constituição brasileira	

- os textos importados do CN deverão adaptar-se ao padrão de catalogação textual do LW, sendo necessária a apresentação das informações exigidas pelo seu cabeçalho, tais como o autor, a data e o local de publicação, o assunto, o meio de distribuição, o tamanho da amostragem, etc. Alguns textos do CN não serão aproveitados porque não possuem essas informações; é o caso dos seguintes:

Teses, Dissertações, Relatórios, etc.	Corpus Técnico-Científico como um todo
Enciclopédias	

- dada a escolha pela amostragem integral dos textos no LW, certos textos do CN estão excluídos do aproveitamento:

Teses e Dissertações	Corpus Didático, especificamente os livros didáticos
Alguns textos do corpus Literário	

Com relação a esse extrato do corpus Didático, pode ser uma alternativa a compilação dos trechos faltantes das obras existentes no CN. De qualquer maneira, esse procedimento implicará, além disso, a obtenção da autorização de uso do material.

- tendo em vista que os textos no LW serão catalogados de acordo com assunto, por exemplo, permitindo a busca automática segundo esse tipo de categoria textual, alguns conjuntos de textos do CN deverão ser reformatados. Nessa operação, o objetivo será individualizar os textos, que passarão a compor arquivos separados. Dentre os conjuntos desse tipo estão os seguintes:

Jornais Alguns textos do corpus literário

Enciclopédias

- a nomeação de todos os arquivos importados do CN deve ser padronizada, permitindo o controle de inserção de textos;

- a fim de evitar falsas impressões e, portanto, análises lingüísticas equivocadas, um programa mais refinado de contagem de ocorrências deve ser desenvolvido. O objetivo perseguido deverá ser o de filtrar do LW aquilo que, de fato, se compreende como "palavra", tornando possível, então, a extração de números precisos sobre os dados do corpus (ex.: número de palavras, número de não-palavras, número de estrangeirismos, números de lemas, etc.)

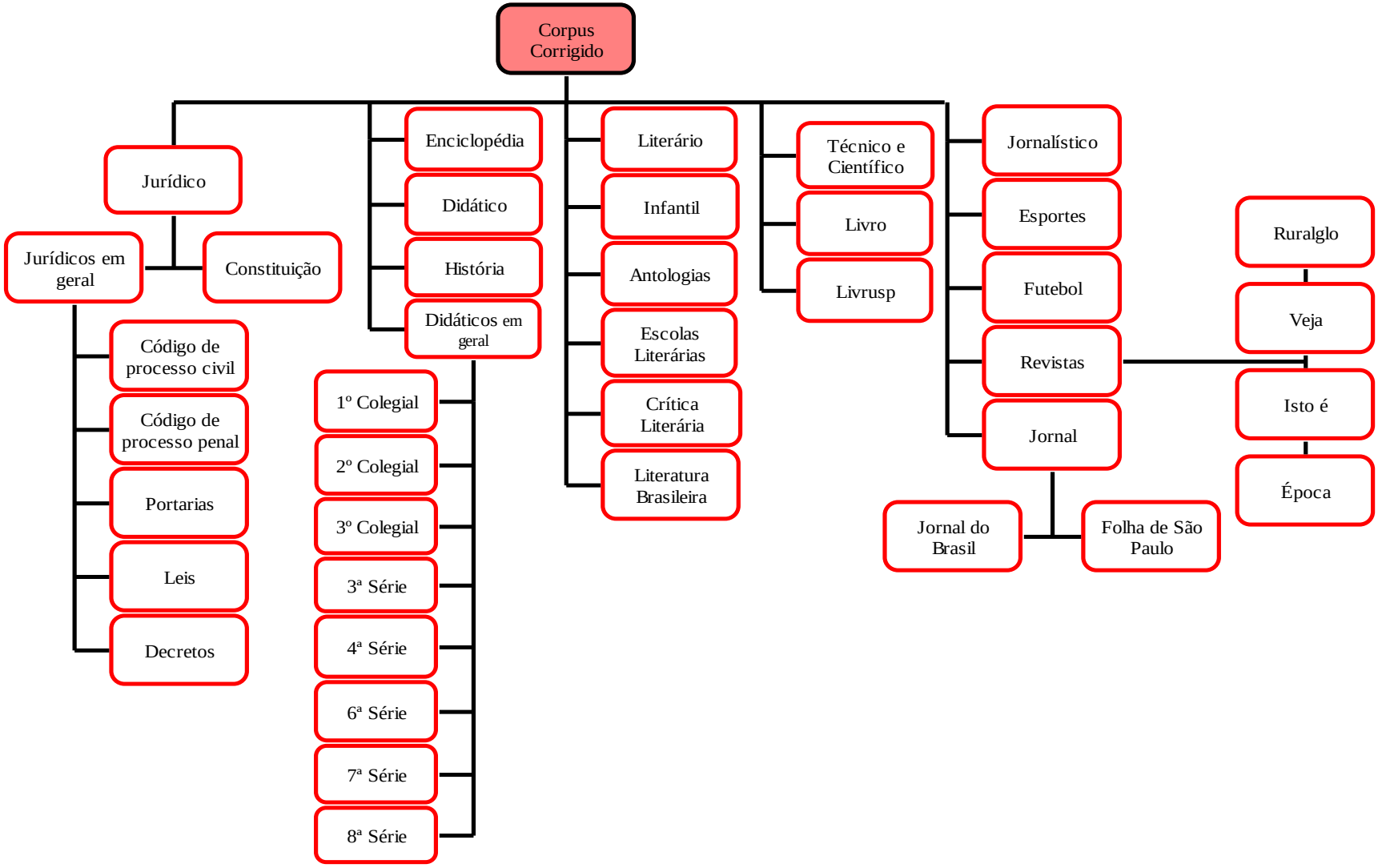
VI. Referências bibliográficas

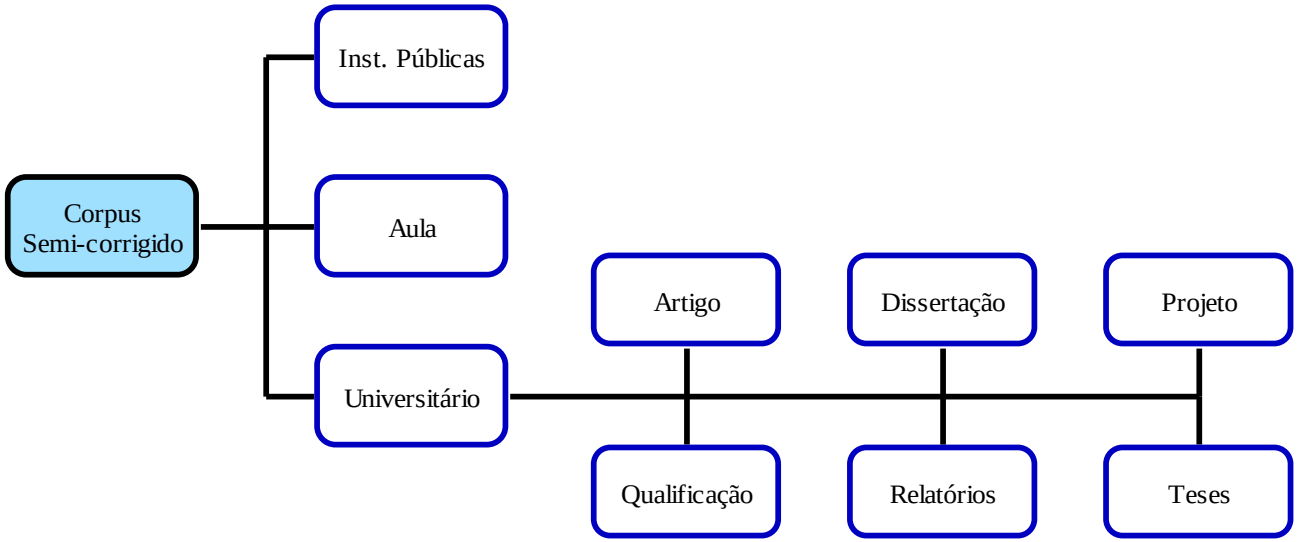
Kuhn, D.; Abarca, E.; Nunes, M.G.V. (2000) *Corpus NILC - Situação em Maio/2000*. (NILC-TR-00-7). Junho, 32p.

Nunes, M.G.V., Hasegawa, R. , Kawamoto, S., Oliveira, M.C.F., Turine, M.A.S., Ghiraldelo, C.M., Oliveira Jr., O.N., Riolfi, C. R., Sikansi, N.S., Martins, T.B.F. (1995) *Desenvolvimento de um revisor gramatical para o português contemporâneo*.

Apêndice I

O Corpus Nilc apresenta uma estrutura de níveis e subníveis de alocação dos textos. Não vamos indicar aqui o nível das pastas, onde os textos propriamente ditos são discriminados. Nos esquemas abaixo, mostramos a organização dos corpora corrigido e semicorrigido divididos em diretórios e subdiretórios.




Corpus
Semi-corrigido
Inst. Públicas
Aula
Universitário
Artigo
Dissertação
Projeto
Qualificação
Relatórios
Teses

Apêndice II

O tamanho dos conjuntos de textos analisados neste documento é dado nas tabelas abaixo. Nelas, constam: a) o número total de ocorrências, por categoria de corpus e/ou textos; b) o número de ocorrências distintas em cada categoria de corpus e/ou textos (ocorrências não-repetidas); e c) o resultado da proporção entre (a) e (b), em cada categoria de corpus e/ou textos. O item (c), especificamente, pretende demonstrar a variabilidade léxica dos corpora. É preciso ter em conta, porém, que nessa contagem os itens não-lingüísticos também estão sendo considerados, conforme discutimos na Seção IV. As tabelas em questão estão devidamente acompanhadas de gráficos, com vistas a permitir uma melhor visualização dos dados numéricos sobre os corpora.

Corpus	No. Ocorrências	Ocorr. não-repetidas	Val. Compar. (em %)
Didáticos	1.147.325	55.766	4,8
Jurídicos	761.852	20.068	2,6
Literários	2.184.620	88.688	4
Téc. & Cien.	1.767.565	72.836	4,1
Jornalísticos	27.203.360	258.082	0,94
Uni			

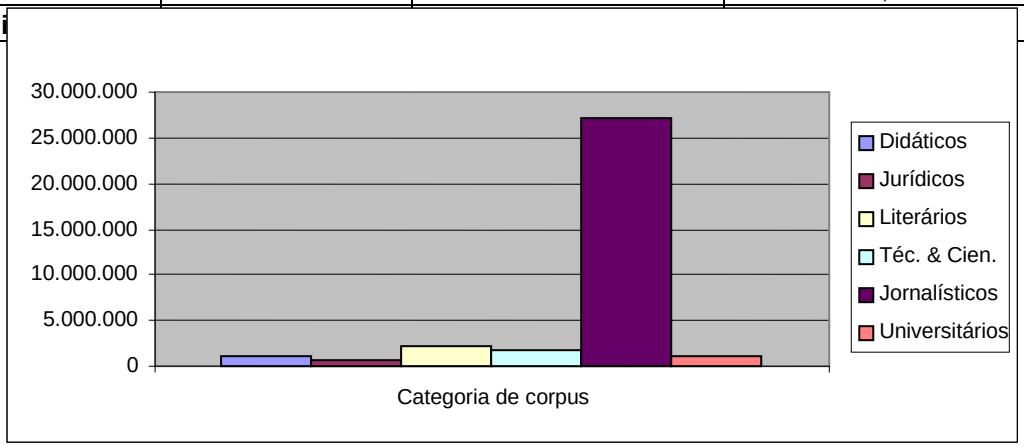


Gráfico 1: Número total de ocorrências nos corpora analisados

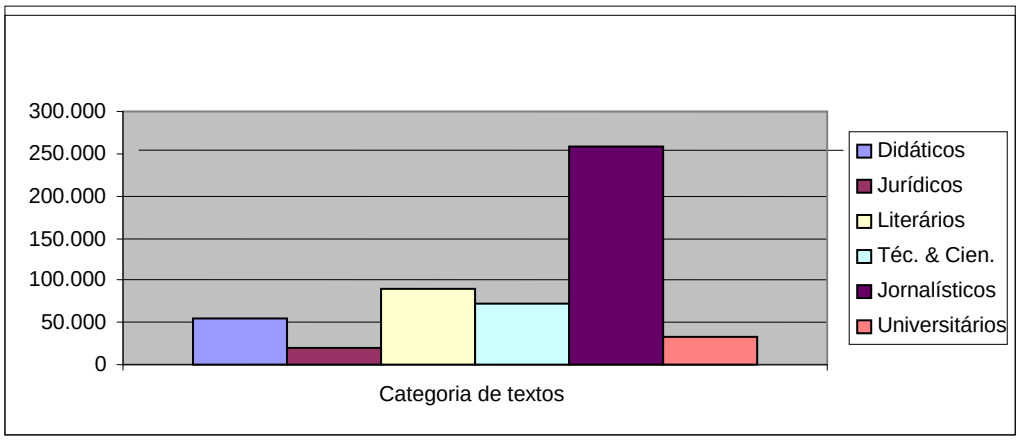


Gráfico 2: Número de ocorrências distintas nos corpora analisados

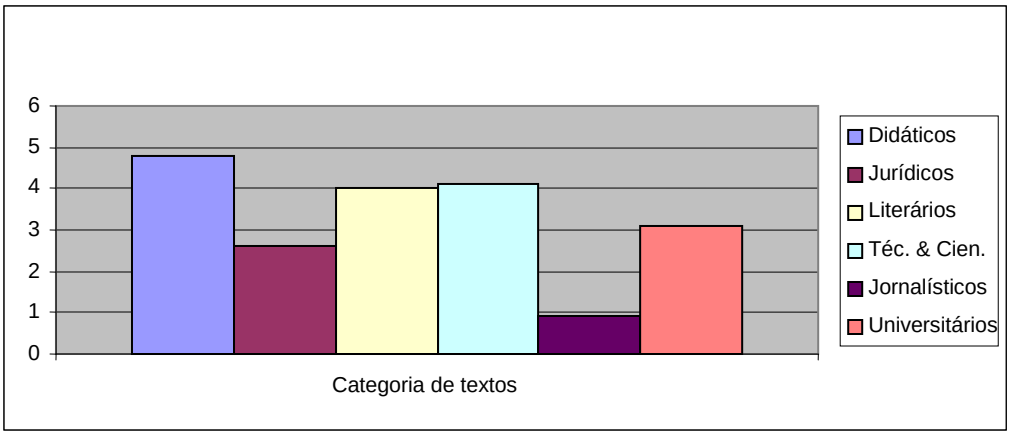


Gráfico 3: Valores comparativos dos corpora analisados

Categoria de textos	No. Ocorrências	Ocorr. não-repetidas	Val. Compar. (em %)
Colégio	758.778	39.262	5,1
Enciclopédia	232.991	26.261	11,2
História	155.556	16.662	10,7

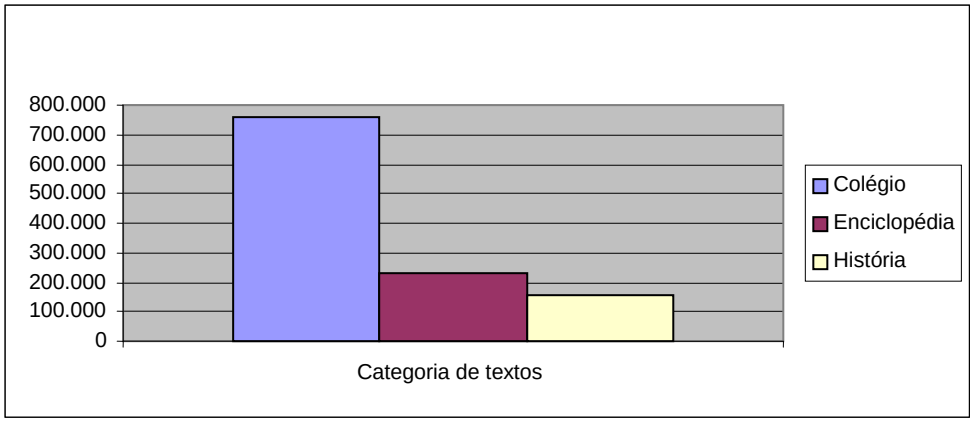


Gráfico 4: Número total de ocorrências no corpus didático

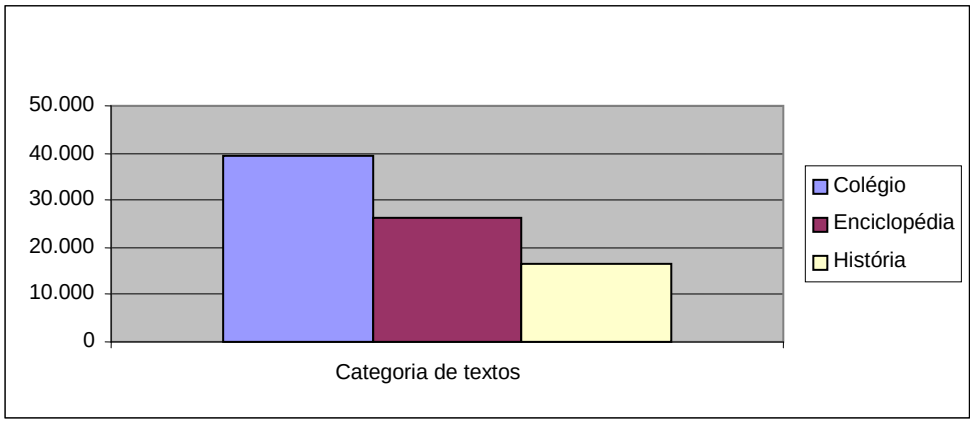


Gráfico 5: Número de ocorrências distintas no corpus didático

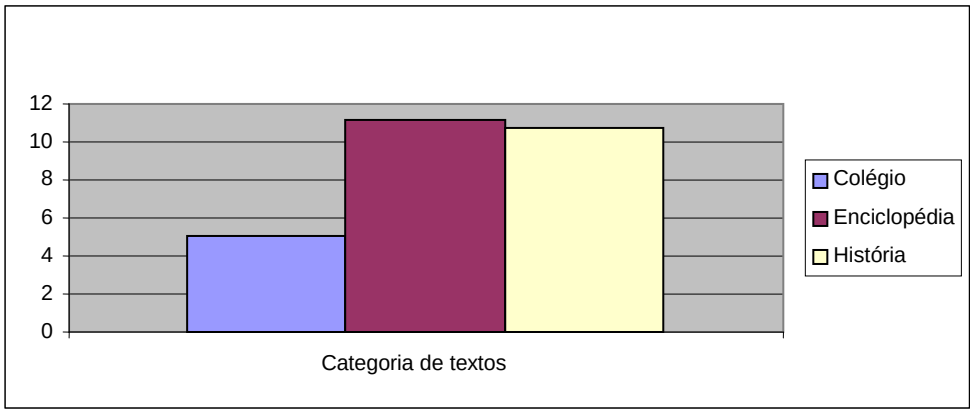


Gráfico 6: Valores comparativos do corpus didático

Categoria de textos	No. Ocorrências	Ocorr. não-repetidas	Val. Compar. (em %)
Cd. Proc. Civil	429.779	13.207	3
Cd. Proc. Penal	221.502	10.207	4,6
Constituição	16.422	2.873	17,5
Decretos	42.406	4.685	11
Leis	45.555	4.587	10
Portarias	6.188	1.471	23,7

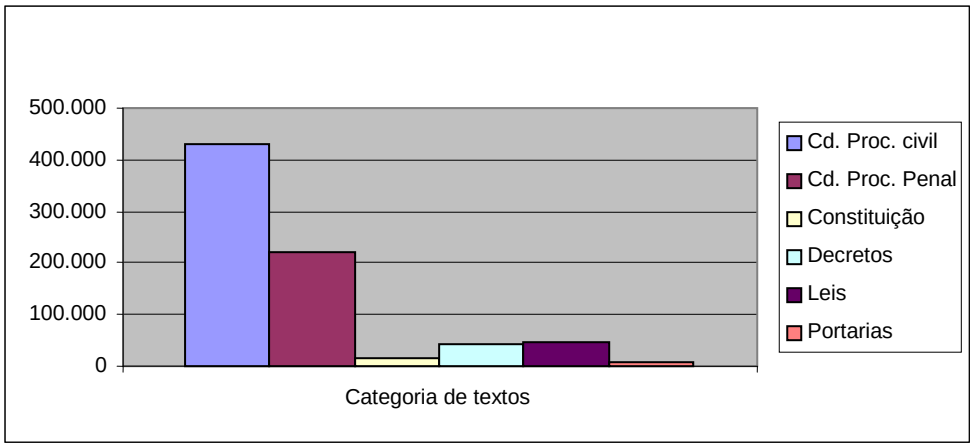


Gráfico 7: Número total de ocorrências no corpus jurídico

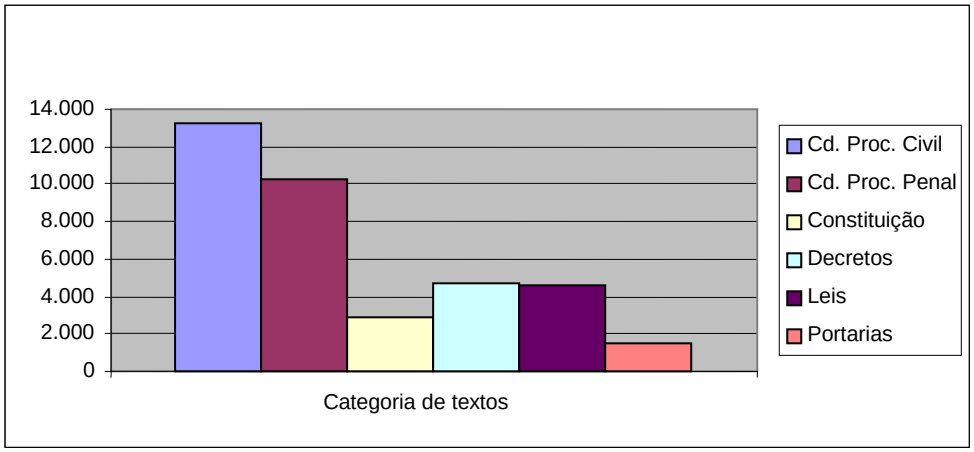


Gráfico 8: Número de ocorrências distintas no corpus jurídico

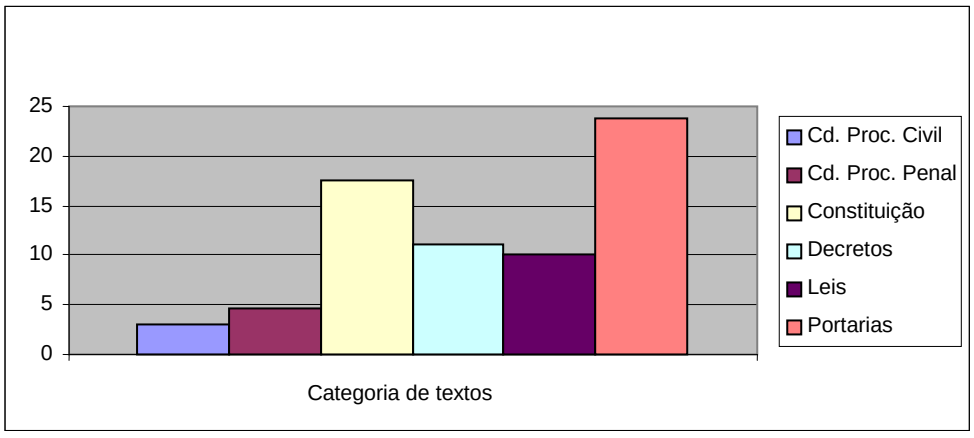


Gráfico 9: Valores comparativos do corpus jurídico

Categoria de textos	No. Ocorrências	Ocorr. não-repetidas	Val. Compar. (em %)
Antologia	51.373	11.042	21,5
Crítica Literária	19.723	4.912	25
Escolas Literárias	11.175	3.223	28,8
Infantil	2.217	881	39,7
Lit. Brasileira	2.056.494	85.641	4,1
Resumos	43.638	8.710	20

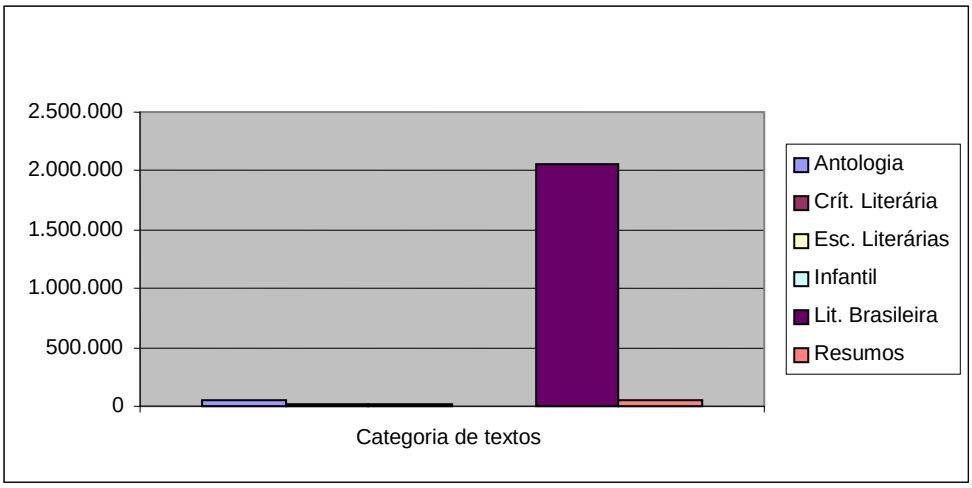


Gráfico 10: Número total de ocorrências no corpus literário

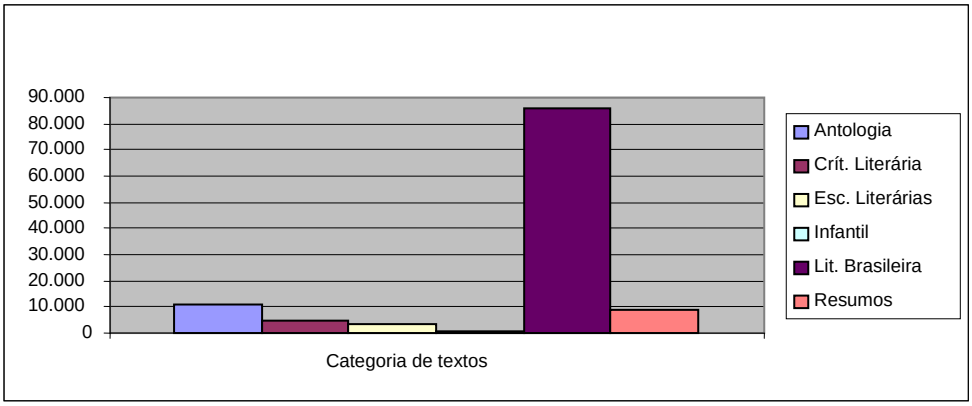


Gráfico 11: Número de ocorrências distintas no corpus literário

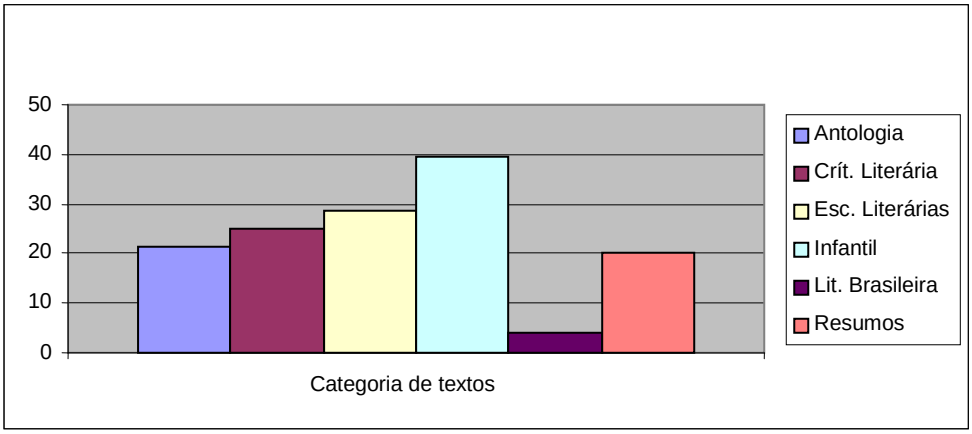


Gráfico 12: Valores comparativos do corpus literário

Categoria de textos	No. Ocorrências	Ocorr. não-repetidas	Val. Compar. (em %)
Esportes	40.368	6.804	16,8
Futebol	96.718	9.755	10
Jornal	26.623.406	255.306	0,96
Revistas	442.868	37.871	8,5

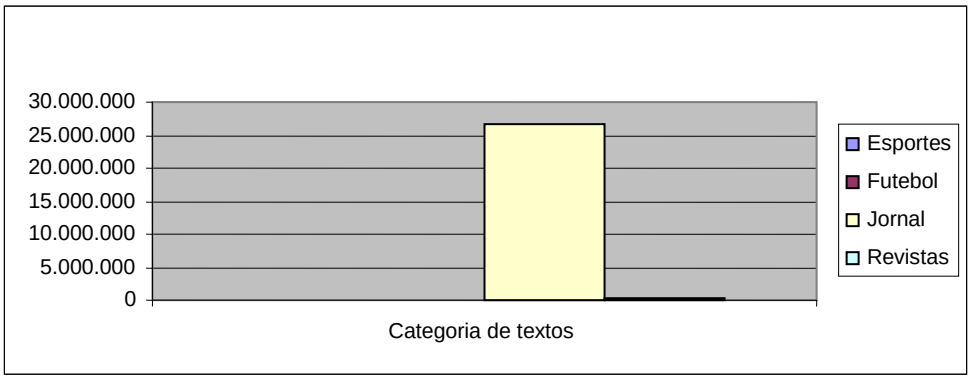


Gráfico 13: Número total de ocorrências no corpus jornalístico

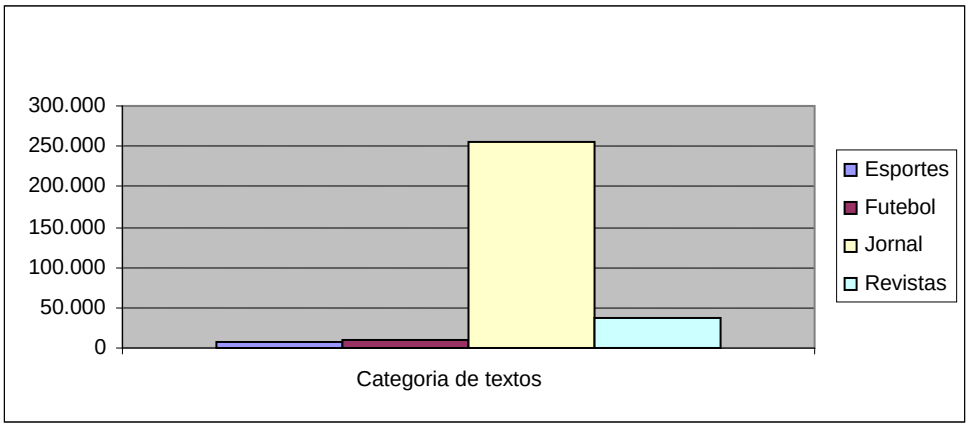


Gráfico 14: Número de ocorrências distintas no corpus jornalístico

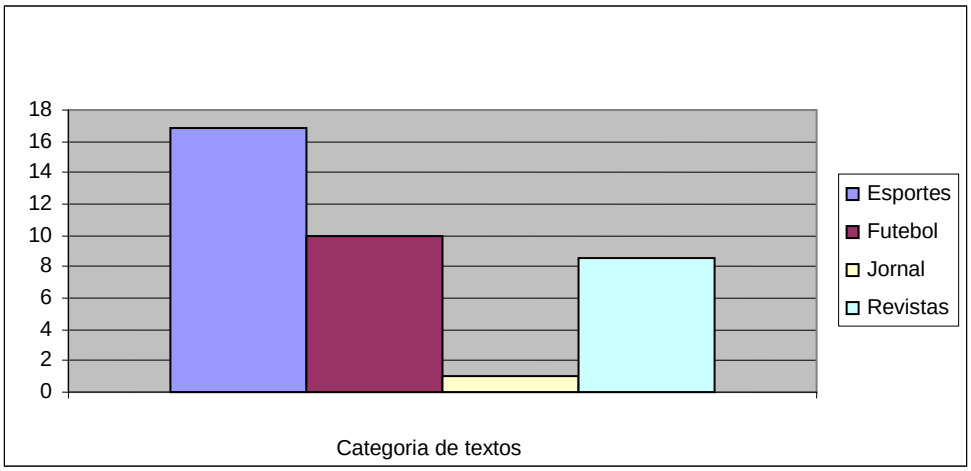


Gráfico 15: Valores comparativos do corpus jornalístico

Categoria de textos	No. Ocorrências	Ocorr. não-repetidas	Val. Compar. (em %)
Livro	173.062	18.235	10,5
Livrusp	1.594.503	68.174	4,2

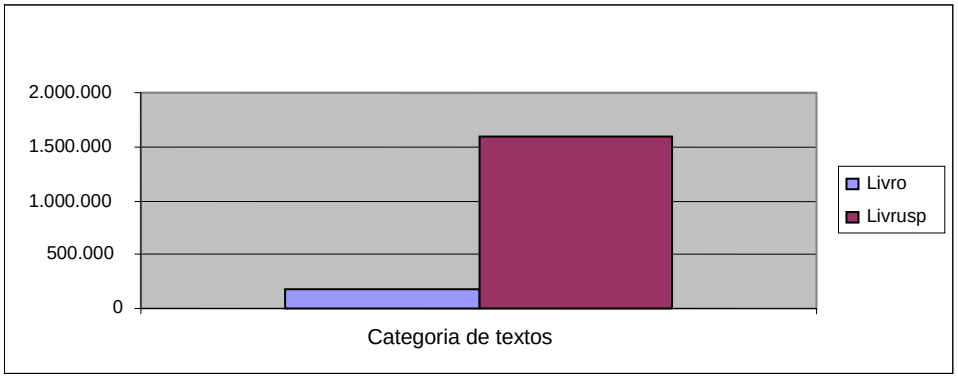


Gráfico 16: Número total de ocorrências do corpus técnico-científico

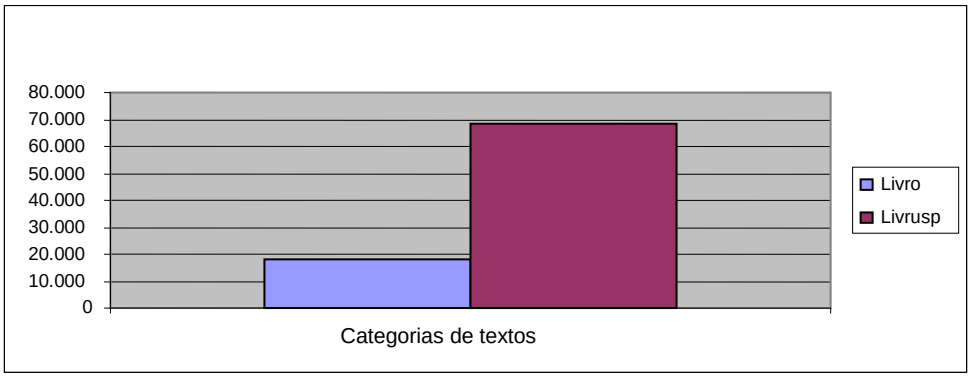


Gráfico 17: Número de ocorrências distintas no corpus técnico-científico

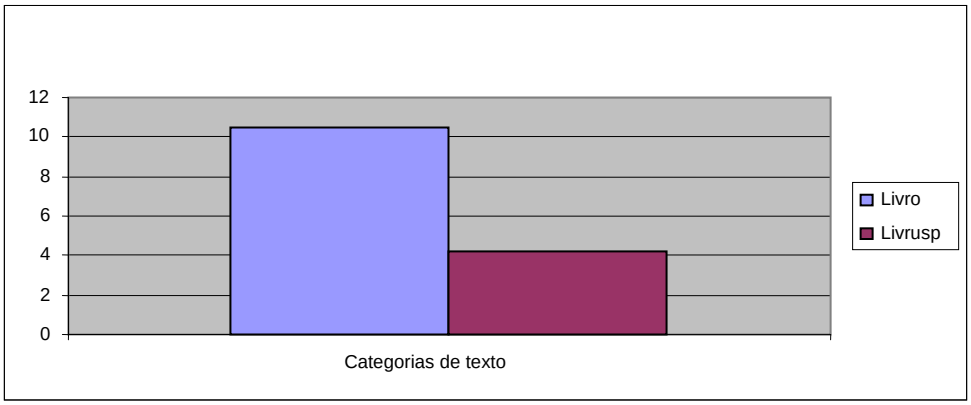


Gráfico 18: Valores comparativos do corpus técnico-científico

Categoria de textos	No. Ocorrências	Ocorr. não-repetidas	Val. Compar. (em %)
Artigo	38.403	5.951	15,5
Qualificação	63.831	5.115	8
Tese	215.347	9.816	4,5
Dissertação	492.623	21.536	4,3
Projeto	33.897	4.374	13
Relatório	183.807	12.111	6,5

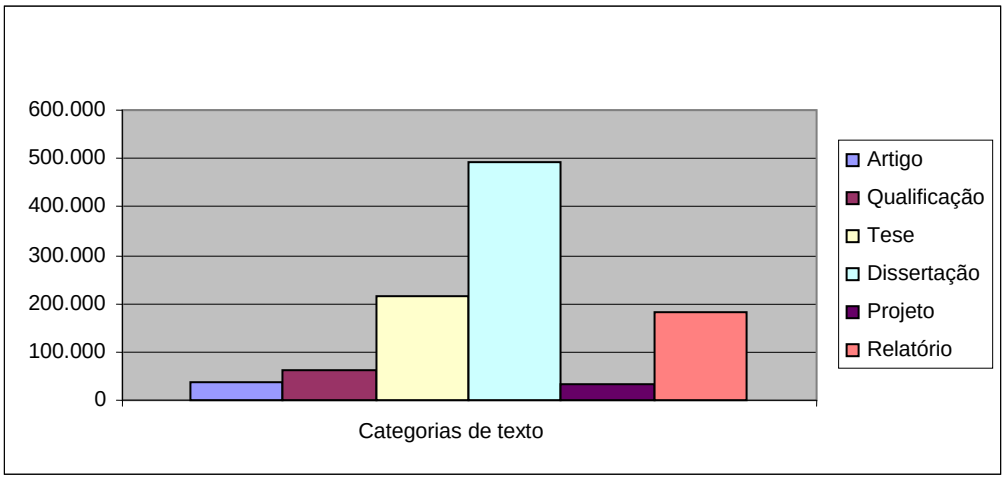


Gráfico 19: Número total de ocorrências no corpus universitário

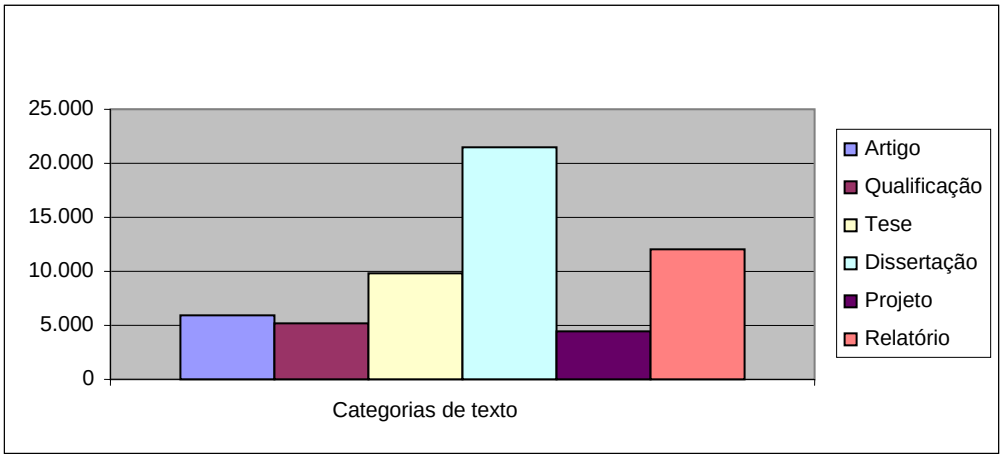


Gráfico 20: Número de ocorrências distintas no corpus universitário

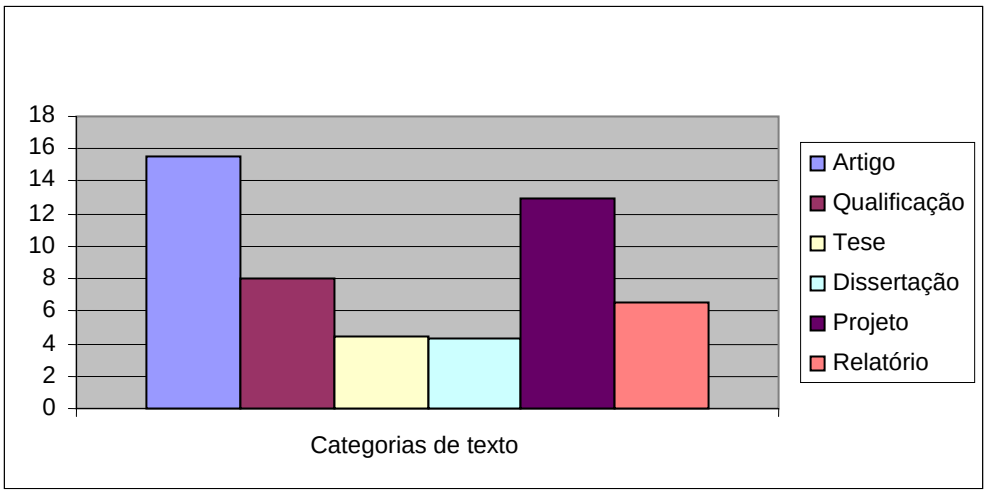


Gráfico 21: Valores comparativos do corpus universitário

Apêndice III

Um programa de contagem automática de ocorrências, bem como de extração da freqüência foi aplicado ao Corpus-Nilc. O resultado deste trabalho foi transformado numa tabela em Access e reproduzido parcialmente aqui.

O conteúdo deste Apêndice é um recorte da tabela em questão. Para compô-lo, privilegiamos: a) os primeiros itens da tabela, i.e., as ocorrências de mais alta freqüência no Corpus; b) itens do meio da tabela; c) itens do final da tabela, i.e., as ocorrências de mais baixa freqüência do Corpus.

A fim de facilitar a visualização dos dados, ordenamos os elementos da tabela: i) por ordem de freqüência; e b) por ordem alfabética, quando o número da freqüência fosse igual. Assim é que surgiu a posição dos itens na contagem. A escolha dos trechos recortados da tabela completa, especialmente quando a quantidade da freqüência era a mesma (cf. dados com freqüência 60, 13, etc.), foi absolutamente aleatória.

Finalmente, cumpre salientar que a contagem utilizada para a construção dessa tabela se refere ao número de ocorrências distintas no CN, i.e., 340.016 ocorrências.

POSIÇÃO	OCORRÊNCIA	FREQÜÊNCIA
1	de	1.701.990
2	a	1.243.051
3	o	1.132.122
4	e	838.584
5	que	782.252
6	do	664.119
7	da	586.378
8	em	477.583
9	para	388.861
10	no	325.882
11	os	321.033
12	um	319.176
13	com	318.696
14	é	313.987
15	não	288.585
16	na	270.495
17	uma	253.416
18	se	208.455
19	as	203.631
20	por	198.214
21	dos	190.678
22	mais	156.462
23	ao	152.496
24	como	144.646
25	são	123.583
26	das	117.751
27	foi	115.022
28	à	114.266
29	página	100.187

30	mas	97.415
31	editoria	97.283
32	ou	88.950
33	ser	87.470
34	sua	82.106
35	ele	79.566
36	pelo	77.443
37	seu	73.322
38	pela	70.140
39	nos	67.044
40	entre	63.451
41	tem	63.379
42	brasil	61.631
43	já	61.268
44	está	61.104
45	sobre	60.602
46	também	59.804
47	paulo	57.990
48	até	57.791
49	segundo	57.059
50	disse	52.192
51	anos	50.238
52	quando	49.337
53	ontem	45.626
54	há	45.621
55	governo	44.405
56	dia	44.311
57	ainda	44.215
58	pode	44.058
59	só	43.719
60	muito	42.643
61	sem	42.600
62	mesmo	42.281
63	era	42.254
64	folha	42.152
65	nas	41.724
66	às	41.327
67	dois	40.821
68	vai	40.821
69	hoje	40.350
70	eu	40.339
71	aos	38.617
72	ano	38.329
73	diz	37.953
74	seus	37.368
75	contra	37.347
76	us	37.138
77	rio	36.889
78	h	35.921
79	presidente	35.878
80	isso	35.309
81	foram	34.754
82	ter	34.479
83	depois	33.730
84	mundo	32.993
85	ela	31.559
86	r	31.515

87	três	30.604
88	local	30.468
89	grande	29.985
90	país	29.756
91	mil	29.557
92	será	29.490
93	onde	29.401
94	todos	28.308
95	estão	28.057
96	tempo	27.888
97	deve	27.485
98	apenas	27.468
99	maior	27.244
100	quem	27.124
101	pessoas	27.053
102	dias	26.431
103	porque	25.817
104	fazer	25.129
105	estado	25.120
106	suas	25.087
107	parte	24.887
108	primeiro	24.792
109	outros	24.763
110	nova	24.545
111	bem	24.139
112	dinheiro	24.073
113	me	23.723
114	cada	23.721
115	especial	23.609
116	caso	23.358
117	reportagem	23.350
118	eles	23.195
119	duas	23.147
120	este	23.142
121	vez	23.103
122	menos	23.001
123	assim	22.806
124	essa	22.355
125	outro	22.292
126	milhões	22.074
127	casa	22.049
128	esse	21.976
129	você	21.785
130	agora	21.741
131	vida	21.115
132	tudo	20.955
133	art	20.752
134	eua	20.717
135	mercado	20.694
136	real	20.639
137	durante	20.503
138	cidade	20.469
139	desde	20.468
140	trabalho	20.318
141	antes	20.297
142	primeira	20.010
143	além	19.972

144	novo	19.904
145	nem	19.714
146	sempre	19.338
147	forma	19.024
148	estava	18.983
149	grupo	18.564
150	melhor	18.509
151	pelos	18.440
152	todo	18.395
153	faz	18.375
154	mês	18.260
155	empresa	18.176
156	nacional	18.168
157	quatro	18.105
158	josé	17.898
159	processo	17.885
160	seja	17.810
161	esta	17.741
162	final	17.628
163	exemplo	17.549
164	preços	17.548
165	qualquer	17.418
166	têm	17.194
167	sistema	17.172
168	carlos	17.053
169	empresas	16.995
170	semana	16.948
171	partir	16.766
172	candidato	16.755
173	plano	16.635
174	programa	16.481
175	tinha	16.426
176	fhc	16.417
177	sendo	16.223
178	fernando	16.076
179	cerca	16.048
180	cotidiano	15.946
181	lei	15.908
182	jogo	15.892
183	acordo	15.795
184	polícia	15.773
185	sul	15.694
186	após	15.694
187	afirmou	15.635
188	alguns	15.627
189	brasileira	15.573
190	podem	15.564
191	outra	15.532
192	banco	15.506
193	esporte	15.502

POSIÇÃO	OCORRÊNCIA	FREQÜÊNCIA
993	disputa	3.800
994	departamento	3.799
995	estudos	3.798
996	sai	3.798

997	comprar	3.795
998	financeiro	3.793
999	casas	3.777
1.000	sistemas	3.773
1.001	francês	3.769
1.002	mar	3.766
1.003	acredita	3.763
1.004	voltou	3.763
1.005	abaixo	3.756
1.006	banda	3.756
1.007	hotel	3.756
1.008	esquerda	3.755
1.009	computador	3.739
1.010	alunos	3.738
1.011	liberdade	3.738
1.012	palavra	3.738
1.013	la	3.735
1.014	acidente	3.733
1.015	execução	3.730
1.016	grau	3.730
1.017	levou	3.729
1.018	médio	3.729
1.019	chamado	3.719
1.020	perto	3.717
1.021	animais	3.709
1.022	discurso	3.709
1.023	julgamento	3.708
1.024	sindicato	3.707
1.025	estilo	3.706
1.026	leva	3.699
1.027	longe	3.691
1.028	tipos	3.686
1.029	bahia	3.684
1.030	comando	3.676
1.031	empresário	3.676
1.032	preso	3.675
1.033	perdeu	3.672
1.034	pequeno	3.667
1.035	dela	3.665
1.036	estas	3.662
1.037	of	3.656
1.038	dúvida	3.653
1.039	parque	3.652
1.040	aumentar	3.651
1.041	títulos	3.651
1.042	stf	3.645
1.043	motivo	3.644
1.044	processos	3.644
1.045	taxas	3.643
1.046	pequena	3.637
1.047	inclusive	3.634
1.048	alemão	3.631
1.049	morreu	3.630
1.050	vontade	3.629
1.051	tempos	3.619
1.052	ricardo	3.618
1.053	saiu	3.616

1.054	importância	3.615
1.055	partes	3.613
1.056	públicos	3.613
1.057	lançamento	3.602
1.058	mário	3.593
1.059	desempenho	3.588
1.060	dado	3.585
1.061	gerais	3.585
1.062	avenida	3.581
1.063	entrou	3.581
1.064	favor	3.578
1.065	vê	3.576
1.066	milhão	3.575
1.067	presente	3.575
1.068	santo	3.571
1.069	significa	3.571
1.070	idéias	3.570
1.071	parreira	3.569
1.072	foto	3.565
1.073	acaba	3.563
1.074	tom	3.557
1.075	podemos	3.555
1.076	estudante	3.546
1.077	exposição	3.540
1.078	pé	3.539
1.079	brizola	3.535
1.080	fundação	3.534
1.081	pediu	3.534
1.082	alimentos	3.529
1.083	andré	3.529
1.084	tendo	3.527
1.085	quarto	3.522
1.086	caixa	3.521
1.087	londres	3.521
1.088	segunda-feira	3.521
1.089	trabalhos	3.521
1.090	meus	3.519
1.091	atualmente	3.517
1.092	medo	3.516
1.093	moda	3.514
1.094	importantes	3.510
1.095	espírito	3.502
1.096	mp	3.499
1.097	veio	3.487
1.098	pm	3.476
1.099	luta	3.471
1.100	times	3.469
1.101	f	3.465
1.102	formas	3.465
1.103	representação	3.465
1.104	próximos	3.453
1.105	senna	3.453
1.106	entidade	3.452
1.107	lima	3.451
1.108	ficaram	3.450
1.109	usp	3.449
1.110	tomar	3.444

1.111	marido	3.439
1.112	procura	3.435
1.113	washington	3.429
1.114	festival	3.425
1.115	jovens	3.425
1.116	mudar	3.419
1.117	sarney	3.419
1.118	ambos	3.418
1.119	objeto	3.411
1.120	acusado	3.410
1.121	debate	3.410
1.122	ganhou	3.409
1.123	metros	3.405
1.124	semanas	3.404
1.125	nove	3.402
1.126	literatura	3.401
1.127	bens	3.395
1.128	projetos	3.394
1.129	fundos	3.388
1.130	participar	3.387
1.131	alves	3.384
1.132	cima	3.383
1.133	mostrar	3.382
1.134	nelson	3.381
1.135	decidiu	3.378
1.136	crítica	3.376
1.137	escritor	3.376
1.138	engenharia	3.373
1.139	pressão	3.371
1.140	números	3.370
1.141	paraná	3.369
1.142	perder	3.369
1.143	podia	3.369
1.144	documento	3.365
1.145	tratamento	3.363
1.146	vi	3.362
1.147	méxico	3.361
1.148	caiu	3.355
1.149	sol	3.354
1.150	recebe	3.349
1.151	seção	3.348
1.152	dizem	3.347
1.153	rússia	3.347
1.154	sangue	3.345
1.155	espanha	3.341
1.156	veja	3.337
1.157	leia	3.334
1.158	israel	3.333
1.159	temporada	3.326
1.160	porta	3.323
1.161	setores	3.322
1.162	tais	3.322
1.163	t	3.311
1.164	dessas	3.310

POSIÇÃO	OCORRÊNCIA	FREQÜÊNCIA

4.687	concorda	750
4.688	deixado	750
4.689	inferno	750
4.690	procuram	750
4.691	tomadas	750
4.692	ampliação	749
4.693	calendário	749
4.694	músculo	749
4.695	perderam	749
4.696	sentimentos	749
4.697	chegado	748
4.698	ênfase	748
4.699	geografia	748
4.700	joel	748
4.701	matriz	748
4.702	irregular	747
4.703	parâmetros	747
4.704	shakespeare	747
4.705	teremos	747
4.706	barreto	746
4.707	encontradas	746
4.708	justificar	746
4.709	apresentaram	745
4.710	cavalos	745
4.711	iniciada	745
4.712	inss	745
4.713	pedaço	745
4.714	baixada	744
4.715	israelenses	744
4.716	perspectivas	744
4.717	caça	743
4.718	correspondência	743
4.719	xuxa	743
4.720	alimento	742
4.721	atrações	742
4.722	chapéu	742
4.723	diria	742
4.724	viúva	742
4.725	algodão	741
4.726	armando	741
4.727	doações	741
4.728	favoráveis	741
4.729	obtenção	741
4.730	propriamente	741
4.731	assassinado	740
4.732	cigarro	740
4.733	exteriores	740
4.734	hábitos	740
4.735	rubião	740
4.736	ocidente	739
4.737	temer	739
4.738	construída	738
4.739	espacial	738
4.740	locação	738
4.741	longos	738
4.742	noruega	738
4.743	revelação	738

4.744	acertar	737
4.745	expressões	737
4.746	irão	737
4.747	pleno	737
4.748	quartas-de-final	737
4.749	servem	737
4.750	visando	737
4.751	barulho	736
4.752	bordo	736
4.753	ensaios	736
4.754	gelo	736
4.755	supermercado	736
4.756	blues	735
4.757	conselheiro	735
4.758	mães	735
4.759	moço	735

POSIÇÃO	OCORRÊNCIA	FREQÜÊNCIA
28.157	capra	60
28.158	cartilagem	60
28.159	casando	60
28.160	casaram	60
28.161	cassa	60
28.162	causadores	60
28.163	celulari	60
28.164	chávez	60
28.165	chegados	60
28.166	cheirar	60
28.167	chesf	60
28.168	cibernética	60
28.169	cirurgiões	60
28.170	city-tour	60
28.171	clarke	60
28.172	coi	60
28.173	colegiado	60
28.174	comandatuba	60
28.175	comemoramos	60
28.176	comparativos	60
28.177	complementam	60
28.178	comportado	60
28.179	compositora	60
28.180	compunham	60
28.181	conformismo	60
28.182	constituições	60
28.183	consultada	60
28.184	contraditória	60
28.185	contraído	60
28.186	contratura	60
28.187	convenientes	60
28.188	convivendo	60
28.189	criançada	60
28.190	cruzaram	60
28.191	culatra	60
28.192	decorreu	60
28.193	dedicou-se	60
28.194	depardieu	60

28.195	desatenção	60
28.196	desativados	60
28.197	desconfiou	60
28.198	desejamos	60
28.199	desmantelamento	60
28.200	despertam	60
28.201	destacou-se	60
28.202	direcionados	60
28.203	discretas	60
28.204	distributivo	60
28.205	diversamente	60
28.206	domingão	60
28.207	doutoramento	60
28.208	eletromiográfico	60
28.209	emirados	60
28.272	lego	60
28.273	levariam	60
28.274	lindsay	60
28.275	lontra	60
28.276	m-	60
28.277	magrão	60
28.278	malone	60
28.279	manguezais	60
28.280	marajás	60
28.281	marchand	60
28.282	marvin	60
28.283	matthaus	60
28.284	mea	60
28.285	metalúrgicas	60
28.286	midi	60
28.287	minimamente	60
28.288	minoritárias	60
28.289	misteriosamente	60
28.290	molar	60
28.291	moleques	60
28.292	monografia	60
28.293	morcegos	60
28.294	mudos	60
28.295	mulatos	60
28.296	narcóticos	60
28.308	ornamentais	60
28.309	ortopedista	60
28.310	pace	60
28.311	palaciano	60
28.312	palmer	60
28.313	peço-lhe	60
28.314	pêndulo	60
28.315	perguntou-me	60
28.316	perplexos	60
28.317	perrin	60
28.318	pescaria	60
28.319	pijama	60
28.320	piquenique	60
28.321	piquetes	60
28.322	pleitos	60
28.323	poliestireno	60
28.324	porta-vozes	60

28.325	postulado	60
28.326	precipício	60
28.327	predileção	60
28.328	profissionalizantes	60
28.329	prole	60
28.330	prolonga	60
28.331	promiscuidade	60
28.332	propositalmente	60

POSIÇÃO	OCORRÊNCIA	FREQÜÊNCIA
66.415	fieira	13
66.416	fielding	13
66.417	fifth	13
66.418	figueiras	13
66.419	filantrópico	13
66.420	filia	13
66.421	filmetes	13
66.422	filos	13
66.423	fiofó	13
66.424	fiorile	13
66.425	firma-se	13
66.426	fiscalizará	13
66.427	fitei	13
66.428	fitoplâncton	13
66.429	fixa-se	13
66.430	fjortoft	13
66.431	flacidez	13
66.432	flamenca	13
66.433	flertes	13
66.434	fletido	13
66.435	florescem	13
66.436	florian	13
66.437	florista	13
66.438	foe	13
66.439	fogosa	13
66.440	fole	13
66.441	folgadamente	13
66.442	folha-de-flandres	13
66.443	foliar	13
66.596	heroicamente	13
66.597	hésio	13
66.598	heterodoxas	13
66.599	hetzel	13
66.600	hicsos	13
66.601	hidrômetros	13
66.602	highland	13
66.603	hilde	13
66.604	hildegard	13
66.605	hipertenso	13
66.606	hipotrofias	13
66.607	hiratsuka	13
66.608	histrionismo	13
66.609	hofmann	13
66.610	holden	13
66.611	holerites	13
67.090	nhã	13

67.091	nhk	13
67.092	nicodemus	13
67.093	nietzscheanos	13
67.094	nietzschiano	13
67.095	nilsen	13
67.096	niveis	13
67.097	no-break	13
67.098	nocautes	13
67.099	nomeie	13
67.100	nordau	13
67.101	norplant	13
67.102	norville	13
67.103	notinha	13
67.104	novacap	13
67.105	noviços	13
67.106	nueva	13
67.107	nuova	13
67.108	nutrasweet	13
67.109	nwa	13

POSIÇÃO	OCORRÊNCIA	FREQÜÊNCIA
150.860	â	2
150.861	àa	2
150.862	aab	2
150.863	aachen	2
150.864	aadequação	2
150.865	aadoção	2
150.866	aafirmação	2
150.867	aage	2
150.868	aalst	2
150.869	a-amilase	2
150.870	aaos	2
150.871	aaplicação	2
150.872	aárea	2
150.873	aaroun-el-raschid	2
150.874	aart	2
150.875	aaury	2
150.876	aavaliação	2
150.877	abá	2
150.878	abades	2
150.879	abafara	2
150.880	abafarma	2
150.881	abafavam	2
150.882	abagail	2
150.883	abaianado	2
150.884	abainhar	2
150.885	abaís	2
150.886	abaité	2
155.946	barulheiras	2
155.947	barulhentamente	2
155.948	barun	2
155.949	basaiev	2
155.950	bascarsija	2
155.951	bascitrus	2
155.952	baseadinho	2
155.953	baseadona	2

155.954	base-aérea	2
155.955	baseando-nos	2
155.956	baseara	2
155.957	basearão	2
155.958	basearem	2
155.959	basearem-se	2
155.960	basea-se	2
155.961	basedos	2
155.962	baseei	2
155.963	baseiem	2
155.964	basendwa	2
155.965	basesestabilizadas	2
155.966	basfond	2
155.967	bas-fonds	2
155.968	basia	2
155.969	básicado	2
155.970	basicamene	2
155.971	basicas	2
155.972	básicasdo	2
155.973	basico	2
155.974	básico-fashion	2
155.975	basidiomiceto	2
155.976	basidiomicetos	2
155.977	basileus	2
167.626	espritualista	2
167.627	espumaterapia	2
167.628	espuminha	2
167.629	espumosa	2
167.630	esquadria	2
167.631	esquadrinhamento	2
167.632	esquadrinha-o	2
167.633	esquadrinharam	2
167.634	esquadrinhou	2
167.635	esquadros	2
167.636	esquartejada	2
167.637	esqueçais	2
167.638	esqueça-nos	2
167.639	esquecei-vos	2
167.640	esquece-os	2
167.641	esqueceríamos	2
167.642	esqueceu-lhe	2
167.643	esquecidinhos	2
167.644	esquema-gigante	2
167.645	esquemaia	2
167.646	esquematizados	2
167.647	esquematizamos	2
167.648	esquematizando	2
167.649	esquentada	2
167.650	esquerda-	2
167.651	esquerda-direita	2
167.652	esquí	2
167.653	esquiaram	2
167.654	esquiável	2
167.655	esquifes	2
167.656	esqui-mó	2
167.657	esquimogenica	2
167.658	esquimogenicas	2

167.659	esquindô	2
167.660	ésquines	2
167.661	esquirol	2
167.662	esquisitíssima	2
167.663	esquisitíssimas	2
167.664	esquisitona	2
167.665	esquivado	2
167.666	esquivam-se	2
167.667	esquivanças	2
167.668	esquizoanálise	2
167.669	esquizoceloma	2
167.670	esquizofrenicamente	2
167.671	esquizofrenico	2
167.672	esquizóides	2
167.673	esquizonte	2
167.674	essarts	2
167.675	essatopologia	2
167.676	essedeve	2
167.677	essencialistas	2
167.678	essencias	2
167.679	essepê	2
167.680	esses-	2
167.681	essetrabalho	2
167.682	essone	2
167.683	essor	2
167.684	essoutra	2
167.685	essurreição	2
167.686	está	2
167.687	estabelcer	2
167.688	estabele	2

POSIÇÃO	OCORRÊNCIA	FREQÜÊNCIA
268.161	infetos	1
268.162	infezlismente	1
268.163	infferno	1
268.164	infibra-permatex	1
268.165	inficionados	1
268.166	infidelity	1
268.167	infieis	1
268.168	infiernos	1
268.169	infiltações	1
268.170	infiltrações-surpresa	1
268.171	infiltrador	1
268.172	infiltrá-las	1
268.173	infiltrará	1
268.174	infiltrarei	1
268.175	infiltra-se	1
268.176	infiltravam	1
268.177	infiltrava-se-lhe	1
268.178	infiltrei	1
268.179	infiltrem	1
268.180	infimamente	1
268.181	infindas	1
268.182	infindavelmente	1
268.183	infini	1
268.184	infinidades	1

268.185	infinítas	1
268.186	infinite	1
268.187	infinitésimal	1
268.188	infinitesimalmente	1
268.189	infinitésimo	1
268.190	infinitivas	1
268.191	infinitivos	1
268.192	infinito-finito	1
268.193	infinitudes	1
268.194	infintesimal	1
268.195	infirmado	1
268.196	infiro	1
268.197	infita	1
268.198	inflaçao	1
268.199	inflaçãoelevada	1
268.200	inflaçãonão	1
322.084	sobretaxe	1
322.085	sobre-tempo	1
322.086	sobretodo	1
322.087	sobretons	1
322.088	sobretudo-	1
322.089	sobretxa	1
322.090	sobreutilização	1
322.091	sobre-utilização	1
322.092	sobreutudo	1
322.093	sobrevalorizando	1
322.094	sobrevalorizar-se	1
322.095	sobrevem	1
322.096	sobrevenientes	1
322.097	sobreviência	1
322.098	sobreviera	1
322.099	sobrevierem	1
322.100	sobrevindos	1
322.101	sobrevirá	1
322.102	sobreviria	1
322.103	sobrevivas	1
322.104	sòbrevivência	1
322.105	sobrevivêncla	1
322.106	sobrevivêneia	1
322.107	so-breviver	1
322.108	sobreviverei	1
322.109	sobreviveremos	1
322.110	sobreviveste	1
322.111	sobrevoá-las	1
322.112	sobrevoá-los	1
322.113	sobrevoarem	1
322.114	sobrevoarmos	1
322.115	sobrevoasse	1
322.116	sobrevoe	1
322.117	sóbrida	1
322.118	sobridade	1
322.119	sobrie-dade	1
322.120	sobriho	1
334.285	utterance	1
334.286	uttilizada	1
334.287	utuando	1
334.288	utulizada	1

334.289	uturuncu	1
334.290	uu	1
334.291	uuêi	1
334.292	uuguai	1
334.293	uum	1
334.294	uuma	1
334.295	uunet	1
334.296	úúnico	1
334.297	úúú	1
334.298	uuuh	1
334.299	üüüüüüüüüüüüüüüü	1
334.300	uvaia	1
334.301	uvaias	1
334.302	uvanobre	1
334.303	uvc	1
334.304	uvdiclo	1
334.305	uversitária	1
334.306	uvfg	1
334.307	uvo	1
334.308	uv-visível	1
334.309	uwajimaya	1
334.310	uwilingiyimana	1
334.311	uwiragiye	1
334.312	uxbridge	1
334.313	uxi	1
334.314	uxoricida	1
334.315	uxoricídio	1
334.316	uyeda	1
334.317	uygur	1
334.318	uygures	1
334.319	uz	1
334.320	uz-	1
334.321	uzal	1
334.322	uzan	1
334.323	uzar	1
334.324	uzcategui	1
334.325	uzcátegui	1
334.326	uzu	1
334.327	uzum	1
334.328	uzumasa	1
334.329	vaalezande	1
334.330	vaan	1
334.331	váarrecadar	1
334.332	vabdelli	1
334.333	vacação	1
334.334	vacaciones	1
334.335	vaca-fashion	1
334.336	vacâncias	1
334.337	vacant	1
334.338	vação	1
334.339	vacareza	1
334.340	vacariou	1
334.341	vacaville	1
334.342	vaccarini	1
334.343	vacchi	1
334.344	vaccinia	1
334.345	vaché	1

334.346	vachek	1
334.347	vachl	1
334.348	vachoe	1
334.349	vacía	1
334.350	vacielei	1
334.351	vaciladas	1
334.352	vacilado	1
334.353	vacilavam	1
334.354	vacilei	1
334.355	vacinacao	1
334.356	vaciná-la	1
334.357	vacinando-os	1
334.358	vacinante	1
334.359	vacinaram	1
334.360	vacinavam	1
334.361	vacínio	1
334.362	vacios	1
334.363	vacom	1
334.364	vácua	1
334.365	vacuolares	1
334.366	vacuolizadas	1
334.367	vadael	1
334.368	vadász	1
334.369	vadeação	1
334.370	vadeando	1
334.371	vadeco	1
334.372	vadée	1
334.373	vade-mecum	1
334.374	vademecums	1
334.375	vades	1