5
Dados coletados

> *one can hardly expect learner corpora to reach the gigantic sizes of native corpora*[91]
>
> (Granger, 1998c, p. 10)

PUC-Rio - Certificação Digital Nº 0610483/CA

Os Capítulos 5 e 6 desta dissertação lidam com os aspectos metodológicos do estudo realizado. Neste caso, são explicados os procedimentos em relação à coleta de dados. Assim sendo, discorre-se a respeito da compilação e formatação dos corpora de pesquisa. Adicionalmente, são descritas as duas populações aqui investigadas: falantes brasileiros de ILE e falantes americanos e britânicos de IL1.

5.1.
Corpora de pesquisa

De forma a estudar a produção escrita em inglês de brasileiros, contrastando-a com a de falantes de IL1, foram utilizados dois corpora (Br-ICLE e LOCNESS, respectivamente), que são descritos nas seções a seguir.

5.1.1.
Br-ICLE

O corpus de pesquisa deste trabalho insere-se em um projeto internacional denominado *International Corpus of Learner English* (ICLE), sediado na Universidade Católica de Louvain e coordenado pela Profa. Dra. Sylviane Granger (http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm). O objetivo do mesmo é compilar subcorpora de redações escritas em língua

[91] Tradução livre para o português: "pode-se dificilmente esperar que corpora de aprendiz alcancem os tamanhos gigantes dos corpora nativos".

inglesa por aprendizes de tal língua oriundos de diferentes nacionalidades. Este projeto é reconhecido como o mais importante exemplo de corpus de ILE (Leech, 1997, p. 20; Granger, 1998, p. 4; Hunston, 2002, p. 15-16).

A compilação possui algumas características comuns como a idade, o contexto de aprendizagem, o nível e o meio das redações (cf. Granger, 1998c, p. 9-10). Em outras palavras, os alunos têm em torno de 20 anos, sendo todos aprendizes avançados de ILE que produzem redações escritas.

Atualmente, 21 subcorpora integram o ICLE, provenientes de diferentes localidades como África do Sul, Alemanha, Brasil, Bulgária, China, Eslovênia, Espanha, Finlândia, França, Grécia, Holanda, Itália, Japão, Lituânia, Noruega, Polônia, Portugal, República Checa, Rússia, Suécia e Turquia. Dentre os 21 mencionados, 11 subcorpora já estão completos, isto é, contêm, no mínimo, 200.000 palavras cada e estão disponíveis comercialmente (Granger, Dagneaux & Meunier, 2002). O corpus totaliza 2,5 milhões de palavras (Baker, Hardie & McEnery, 2006, p. 93).

O ramo brasileiro do ICLE, denominado Br-ICLE (acrônimo para *The Brazilian Portuguese Sub-corpus of ICLE*), iniciado em 29 de abril de 1999 (Berber Sardinha, 2001), é coordenado pelos Profs. Drs. Tony Berber Sardinha e Stella Esther Ortweiler Tagnin e sediado na Pontifícia Universidade Católica de São Paulo (PUC-SP). É através do site oficial do Br-ICLE (http://www2.lael.pucsp.br/corpora/bricle/index.htm) que os colaboradores podem dar entrada nos dados coletados através de um sistema protegido por senha. Todas as informações inseridas no sistema (cf. Anexo 10.1) são registradas e armazenadas com um número específico, facilitando assim o acesso às mesmas em um estágio posterior (cf. Anexo 10.2). Apesar de haver redações da PUC-SP, UFSC (Universidade Federal de Santa Catarina) e USP (Universidade de São Paulo) no estágio atual do Br-ICLE, as mesmas não foram utilizadas neste estudo principalmente porque os dados não estão disponíveis na íntegra na Internet. Assim sendo, não há como verificar em que condições as coletas foram realizadas nem com quais participantes. Foram, então, somente consideradas para fins de análise as redações coletadas no Estado do Rio de Janeiro.

PUC-Rio - Certificação Digital Nº 0610483/CA

5.1.1.1.
Princípios de coleta de dados

De acordo com os critérios de compilação do Br-ICLE, os textos a serem inseridos no corpus podem ser de duas naturezas distintas: literária e argumentativa. As redações coletadas para esta pesquisa foram somente de natureza argumentativa de forma que se pudesse controlar o corpus a ser investigado.

Todos os textos foram produzidos em língua inglesa. Como a coleta foi realizada no Brasil, país no qual esta língua não é oficial, caracteriza-se a mesma como LE. Assim sendo, o estudo de características lingüísticas presentes nas redações dos sujeitos de pesquisa constitui-se, na verdade, em um estudo da escrita de brasileiros em ILE.

Para contribuir com a pesquisa, os alunos deveriam ser caracterizados como de nível avançado. No âmbito do ICLE, estes aprendizes são identificados como alunos universitários no terceiro ou quarto ano de estudo do curso de Letras ou correspondentes. Como o sistema educacional de várias universidades brasileiras baseia-se em créditos cursados semestralmente, adaptou-se o modelo em tela aos padrões locais. Em outras palavras, considerou-se como aluno avançado aquele que estivesse no quinto período letivo, o que corresponde aproximadamente à metade dos semestres / créditos do curso de Letras.

Os temas fornecidos aos alunos são os que constam em uma lista compilada pelos responsáveis pelo ICLE. Tal lista contempla quatorze possibilidades, que são apresentadas a seguir.

(A) *Crime does not pay.*
(B) *The prison system is outdated. No civilised society should punish its criminals: it should rehabilitate them.*
(C) *Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.*
(D) *A man / woman's financial reward should be commensurated with their contribution to the society they live in.*
(E) *The role of censorship in Western society.*
(F) *Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.*
(G) *All armies should consist entirely of professional soldiers: there is no value in a system of military service.*
(H) *The Gulf War has shown us that it is still a great thing to fight for one's country.*
(I) *Feminists have done more harm to the cause of women than good.*

PUC-Rio - Certificação Digital Nº 0610483/CA

(J) *In his novel Animal Farm, George Orwell wrote "All men are equal: but some are more equal than others." How true is this today?*
(L) *In the words of the old song "Money is the root of all evil."*
(M) *Europe: birth of a nation or loss of sovereignty?*
(N) *In the 19th century, Victor Hugo said: "How sad it is to think that nature is calling out but humanity refuses to pay heed." Do you think it is still true nowadays?*
(O) *Some people say that in our modern world, dominated by science technology and industrialisation, there is no longer a place for dreaming and imagination. What is your opinion?*

Durante a realização da tarefa escrita, os alunos puderam consultar materiais de referência se eles assim o desejassem ou se tal prática fosse permitida por seus professores. Tais recursos são entendidos como dicionários (monolíngües ou bilíngües), gramáticas, páginas da Internet, livros diversos, entre outros. Contudo, os alunos foram solicitados a não copiar partes de artigos ou livros sem indicar a devida referência assim como também não deveriam ter ajuda de terceiros.[92]

Em relação ao tamanho, solicitou-se que os alunos escrevessem textos de, no mínimo, 500 palavras. O número máximo de palavras permitido em cada redação era de 1.000 palavras. Como cada aluno só pôde contribuir com uma única redação, tem-se que a colaboração por aluno também variou entre 500 e 1.000 palavras. A restrição da quantidade de textos por participante é especialmente importante de forma a evitar que a análise realizada fosse baseada em usos idiossincráticos da língua inglesa.

Além da redação propriamente dita, os alunos preencheram um questionário (cf. Anexo 10.3) no qual eles registraram (a) características da redação, (b) condições de realização da mesma, (c) informações pessoais e familiares, (d) formação educacional e utilização / exposição a línguas estrangeiras, especialmente o inglês, e (e) autorização para que a redação pudesse ser utilizada para fins científicos. Assim sendo, na primeira parte do questionário, os alunos indicaram o título da redação e o seu tamanho aproximado. Em um segundo momento, os alunos foram solicitados a informar se houve um limite de tempo para a realização da tarefa, se a mesma foi parte integrante da avaliação de um curso e se eles utilizaram materiais de referência. Posteriormente, questionava-se a respeito de nome, idade, sexo, nacionalidade, língua materna – do próprio

[92] Não há, entretanto, como afirmar que a orientação foi seguida no tocante às redações que não foram escritas em sala de aula.

PUC-Rio - Certificação Digital Nº 0610483/CA

participante e de seus pais – e línguas faladas em casa. Na quarta seção do questionário, os alunos indicaram a língua utilizada tanto no ensino fundamental como no ensino médio, o curso atual deles na universidade, o ano de estudo, a instituição e a língua de instrução na mesma. Além disto, era necessário informar o número de anos de estudo de inglês (na escola e na universidade), eventuais estadias em países cuja primeira língua é o inglês, e o domínio de outras línguas estrangeiras em ordem decrescente de proficiência. Finalmente, os alunos foram solicitados a datar e assinar o questionário, autorizando o uso de suas redações em pesquisas científicas (ver Anexo 10.3).

O preenchimento deste questionário permite a realização de um grande número de pesquisas posteriores a respeito da influência dos fatores mapeados no mesmo. Pode-se verificar, por exemplo, se há alguma variação na escrita dos alunos investigados de acordo com o número de anos de estudo de inglês e/ou a língua materna dos alunos, entre outras possibilidades.

PUC-Rio - Certificação Digital Nº 0610483/CA

5.1.1.2. Coletas de dados

Foram realizadas coletas de dados, seguindo os critérios do ICLE, em quatro universidades no Estado do Rio de Janeiro com vistas à compilação do corpus de estudo. Os locais de coleta foram selecionados uma vez que os mesmos são centros de ensino de qualidade reconhecidos pelas autoridades governamentais brasileiras como apontam as avaliações destes cursos (Provão e Enade) realizadas pelo Ministério da Educação e Cultura. Foram envolvidas nas coletas de dados uma universidade particular e três universidades públicas: Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Universidade do Estado do Rio de Janeiro (UERJ), Universidade Federal Fluminense (UFF) e Universidade Federal do Rio de Janeiro (UFRJ).

O número de participantes da pesquisa por cada centro universitário apresenta variação como indica a Figura 11.

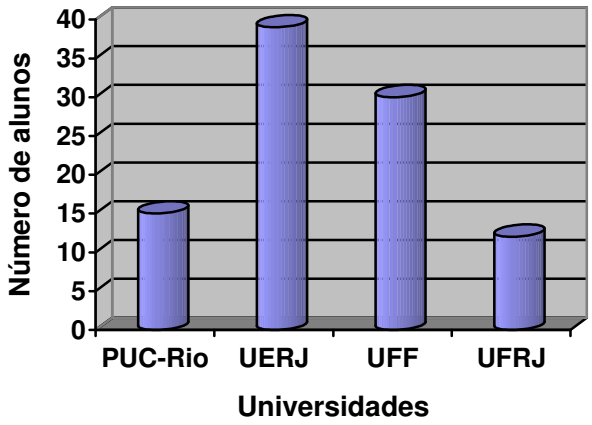

Figura 11: Distribuição de alunos por universidade (Br-ICLE)

A maior concentração de alunos é proveniente da UERJ, que totaliza 39 alunos participantes. Em segundo lugar, há a colaboração de 30 alunos da UFF. A participação de alunos da PUC-Rio e da UFRJ parece ser, de certa forma, semelhante com 15 e 12 alunos, respectivamente. Ao todo, 96 alunos nestas quatro universidades distintas contribuíram para esta pesquisa.

Na PUC-Rio,[93] foi realizada uma coleta com os alunos da disciplina Discurso Escrito II, oferecida pelo Departamento de Letras para o Bacharelado e a Licenciatura em Português, Inglês e Literaturas correspondentes (BPI e LPI), e para o Bacharelado em Tradução – Português e Inglês (BTI). A outra coleta foi realizada com alunos de Lingüística Aplicada: Língua Estrangeira, igualmente oferecida pelo Departamento de Letras, mas somente para o curso LPI. As coletas foram realizadas em um dia de aula, com a presença do pesquisador e com a duração média de duas horas para a realização da tarefa escrita. No entanto, apesar do tempo dado, alguns alunos não conseguiram terminar a redação. Nestes casos específicos, os alunos foram solicitados a completá-las em casa. Em uma das turmas, os alunos fizeram as redações no laboratório de computadores, o que

[93] No caso da PUC-Rio, há uma diferença em relação às outras universidades aqui investigadas: os alunos são nivelados ao entrarem na universidade. Isto significa que um ingressante não necessariamente começa a cursar as disciplinas específicas da área de língua e/ou literatura inglesa no ano em que entra na universidade. Neste caso específico, os alunos realizam um teste de nivelamento. O resultado pode indicar que o aluno tenha que cursar alguns dos seis períodos de inglês que são oferecidos antes de o mesmo começar a estudar as disciplinas curriculares de inglês. Caso o aluno tenha uma base sólida, ele poderá ser autorizado a cursar logo a primeira matéria em língua inglesa da grade de BPI, LPI e BTI, a saber, Língua Inglesa: Texto e Leitura I. Nesta pesquisa, para fins de contabilização do período no qual o aluno se situava, foram considerados tão somente os períodos previstos no fluxograma oficial dos três cursos citados.

PUC-Rio - Certificação Digital Nº 0610483/CA

permitiu que as mesmas fossem entregues em formato digital. Na outra turma, os alunos fizeram a redação em papel e foram, posteriormente, digitadas pelo pesquisador.

A coleta na UERJ foi realizada com alunos de Língua Inglesa VI e Língua Inglesa VIII do Bacharelado e/ou Licenciatura em Inglês e respectivas Literaturas no primeiro semestre de 2005. As redações coletadas junto aos alunos do sexto período fizeram parte da avaliação dos mesmos na disciplina de inglês. Contudo, tais correções não interferiram na coleta de dados visto que esta foi realizada antes da correção ser realizada. As redações foram entregues à professora da turma em formato digital por meio da Internet, mais especificamente por correio eletrônico. No caso dos alunos do oitavo período, a participação foi voluntária no sentido de que a tarefa não fez parte de suas avaliações e as redações não foram realizadas durante uma aula. Os textos foram redigidos em casa, sem tempo determinado e com livre acesso a materiais de referência assim como aceito pelos critérios do ICLE. À semelhança da coleta feita no sexto período, todas as redações foram entregues em formato digital, mas neste caso a coleta foi feita pelo pesquisador.

Uma nova coleta na UERJ foi realizada no segundo semestre de 2006 pelo pesquisador. Desta vez, alunos cursando Língua Inglesa VII participaram da pesquisa. As redações não fizeram parte da avaliação dos alunos, mas foram realizadas em sala de aula durante dois tempos, ou seja, uma hora e quarenta minutos. Os alunos puderam consultar materiais de referência, mas muitos deles não o fizeram por não os terem à mão. As redações neste caso foram feitas manualmente e entregues em papel para que pudessem ser digitadas posteriormente pelo pesquisador.

Na UFF, a coleta foi realizada durante aulas de língua inglesa; porém, grande parte das redações não foi utilizada como instrumento de avaliação por parte do professor da turma. De uma forma geral, os alunos cursavam o quarto ou o quinto ano da Licenciatura em Português e Inglês. As redações foram majoritariamente escritas manualmente e tiveram que ser digitadas posteriormente pelo pesquisador.

A coleta realizada na UFRJ foi informal, ou seja, solicitou-se que alguns alunos escrevessem as redações para a pesquisa. Os alunos que contribuíram para a pesquisa, integrantes do grupo de pesquisa REDES (*Research and Development in Empirical Studies*), fizeram as redações em suas casas, não tiveram limite de

PUC-Rio - Certificação Digital Nº 0610483/CA

tempo, puderam consultar quaisquer materiais que desejassem e entregaram seus textos em formato digital por correio eletrônico.

No primeiro semestre de 2007, uma nova coleta foi realizada na UFRJ. Diferentemente da anterior, esta foi realizada com estagiários do Colégio de Aplicação da UFRJ (CAp-UFRJ), que cursavam o quinto ano de Letras, habilitação Português e Inglês. As condições de realização destas redações foram semelhantes às descritas anteriormente. A única diferença relacionou-se ao fato de que a escritura das redações contou como atividade de estágio supervisionado para os licenciandos que aceitaram participar da pesquisa.

5.1.1.3. Condições de realização das redações

PUC-Rio - Certificação Digital Nº 0610483/CA

É possível verificar as condições de realização das redações a partir da análise dos questionários preenchidos pelos alunos quando da coleta de dados. Em termos da questão temporal, nota-se que a grande maioria dos textos foi produzida sem nenhum tipo de limite. Uma análise detalhada indica que 77,08% dos participantes puderam escrever suas redações no tempo que lhes conviesse enquanto somente 22,92% dos alunos tiveram que escrever os mesmos em um período determinado. Neste último caso, os alunos tiveram que realizar a tarefa em sala durante suas aulas.

Quanto à redação ter sido utilizada como instrumento de avaliação, nota-se que este não foi o padrão das coletas realizadas. Apenas 25,00% das redações foram consideradas como provas e avaliadas pelos professores das disciplinas nas quais as coletas foram realizadas. De forma contrária, a maioria – 75,00% – foi escrita somente para fins de pesquisa científica.

Comenta-se também a respeito da utilização de materiais de referência pelos participantes da pesquisa. A maioria dos alunos (62,50%) recorreu a tais materiais enquanto 37,50% dos mesmos não o fizeram. Este panorama, no entanto, não deve ser interpretado como uma escolha livre dos alunos. A consulta neste caso está diretamente relacionada às condições de realização da redação (isto é, em casa ou em sala de aula) e ao conhecimento prévio da realização da tarefa em sala de aula (isto é, se os alunos foram informados antecipadamente que escreveriam o texto em sala).

5.1.1.4.
Formatação do corpus

Após a coleta dos dados, as redações entregues em papel e escritas à mão foram digitadas. Durante esta etapa, tomou-se cuidado para que a produção textual não fosse descaracterizada seja pela correção de desvios cometidos pelos alunos ou pela inserção de desvios não existentes nos textos originais. Optou-se por um trabalho em três etapas de forma a diminuir a possibilidade de erro humano. Primeiramente, as redações foram digitadas. Posteriormente, os arquivos finais foram impressos de forma que pudessem ser conferidos manualmente com os originais entregues pelos alunos. Finalmente, as eventuais incorreções cometidas pelo pesquisador foram corrigidas diretamente nos arquivos armazenados no computador.

Alguns procedimentos específicos tiveram que ser seguidos tendo em vista os critérios de compilação estabelecidos pelo ICLE. Retirou-se, então, toda e qualquer referência aos nomes dos alunos, aos títulos das redações e às datas de realização dos mesmos dos textos a serem inseridos no corpus. As redações, após a formatação, assumiram a forma ilustrada pela Figura 12.

PUC-Rio - Certificação Digital Nº 0610483/CA

Facing the current situation at University, scholars may conclude that there is a huge complication in the Brazilian educational system. First, most of students come from the elementary school with no basis because it does not prepare them practical and intellectually for the University. Second, a large number of Brazilian families based in the cities, and mainly in the countrified areas, suffer from hunger.

The Brazilian Government does not spend enough money on education and the amount being spent is used ineffectively and is badly distributed. Therefore, students are commonly underprepared for a practical and academical tuition that is usually applied at Universities. Furthermore, the high unemployment rates, hunger and public transportation crisis leads students to leave the school in order to work and to support their families.

In Brazil, there is the belief that only those who are graduate would have the opportunity to get a job which will offer them the best salaries. In fact, those who have no degree may undergo much more difficulties such as to live without health plan, to tolerate hours stuck in the traffic just to reach the place of work. Graduates, by contrast, have more favourable set of circumstances for having a better standard of living. All this demand for degrees has decrease the quality of graduate education, although the Brazilian Ministry of Education maintains an elaborate and well-reputed system of peer review evaluation for the 1,293

programs half of which doctoral degrees to control the standard of education in Universities.

Finally, Universities has been in crisis due to a troublesome administration in Brazil. The Brazilian's Government lack of commitment with their citizens has reduced the value of Universities degrees and it has not been taken seriously by employers because the academic rank conferred by university after the conclusion of a course at University is not creditable to consider that the graduate student is a competent professional.

Figura 12: Exemplo de redação em formato digital (091P4cSc)

Palavras ou expressões ilegíveis foram substituídas por <?>. Referências bibliográficas no corpo da redação no formato 'autor, ano de publicação' ou 'autor, ano de publicação, número da página' foram substituídas pelo símbolo <R>. Citações foram excluídas e substituídas por <*>. A única exceção a esta última regra consiste em citações curtas e/ou integradas às sentenças. Assim sendo, um trecho como o reproduzido na Figura 13 teve que ser alterado para se adequar às normas do corpus, conforme ilustrado na Figura 14.

"The world is a vampire." This line was said by the American singer and songwriter Billy Corgan in one of his songs with his band The Smashing Pumpkins. Many times, this is the exact feeling that people have in relation to the modern world in which we are living nowadays. It seems that our world sucks our time, our pleasures, our capacity to think and analyse information since it all comes in such a fast way. And that is the reason that makes so many people wonder where the dreams fit in our rushed lives.

Figura 13: Parágrafo original (041E4oSn)

<*> This line was said by the American singer and songwriter Billy Corgan in one of his songs with his band The Smashing Pumpkins. Many times, this is the exact feeling that people have in relation to the modern world in which we are living nowadays. It seems that our world sucks our time, our pleasures, our capacity to think and analyse information since it all comes in such a fast way. And that is the reason that makes so many people wonder where the dreams fit in our rushed lives.

Figura 14: Parágrafo após formatação (041E4oSn)

A citação inicial que inicia o parágrafo da redação original foi suprimida na versão final e substituída pelo símbolo <*>, que é automaticamente ignorado pelo programa computacional utilizado nesta pesquisa.

Todos os arquivos foram salvos em formato TXT (texto sem formatação) de forma que pudessem ser lidos adequadamente pelo programa adotado. Ademais,

PUC-Rio - Certificação Digital Nº 0610483/CA

os textos também tiveram que ser tratados para que os números de parágrafos e de frases em cada um das redações fossem corretamente levantados. Assim sendo, inseriu-se um espaço após a última frase de cada parágrafo e duas quebras de linha após o último parágrafo de cada redação. Finalmente, também foram convertidas as aspas curvadas em aspas retas por causa do programa empregado (cf. Seção 6.1).

A nomeação dos arquivos foi a última etapa do tratamento de dados. Cada arquivo recebeu um nome de oito dígitos (limite imposto pela versão do programa utilizado) com base em certas características da redação com vistas à identificação da origem e condições de realização da redação a partir da leitura do nome da mesma. A nomeação seguiu a seguinte lógica:

(a) seqüência de três dígitos: indicação do número da redação;

(b) uma letra maiúscula: indicação da universidade à qual o aluno pertence, a saber, 'E' para UERJ, 'F' para UFF, 'P' para PUC-Rio e 'R' para UFRJ;

(c) um número: indicação do período do aluno, sendo '3', '4' ou '5' as únicas possibilidades;

(d) uma letra minúscula: indicação do tema da redação (cf. Seção 5.1.1.1), sendo as quatorze primeiras letras do alfabeto ('a' – 'o') as possibilidades neste caso;

(e) a letra 'S' ou 'N': indicação do uso ('S') ou não ('N') de materiais de referência para a realização da redação;

(f) a letra 'n' ou 'c' minúsculas: indicação da marcação ('c') ou não ('n') do tempo para a escritura da redação.

Desta forma, o exemplo dado na Figura 12 que tem o código 091P4cSc como nome do arquivo TXT refere-se ao 91º texto do corpus, produzido por um aluno da PUC-Rio no 4º período. Esta redação teve como base o tema 'c' (*Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value*). O aluno fez uso de materiais de referência quando da produção do texto, que, por sua vez, teve seu tempo cronometrado.

5.1.1.5. Caracterização do corpus

O corpus de pesquisa contém 51.430 palavras e 5.455 palavras diferentes em 96 redações (cf. Anexos 10.4 – 10.8). Considerando que os dados foram

PUC-Rio - Certificação Digital Nº 0610483/CA

coletados em quatro universidades diferentes, pode-se verificar a contribuição de cada universidade em relação ao número de palavras conforme ilustra a Figura 15.

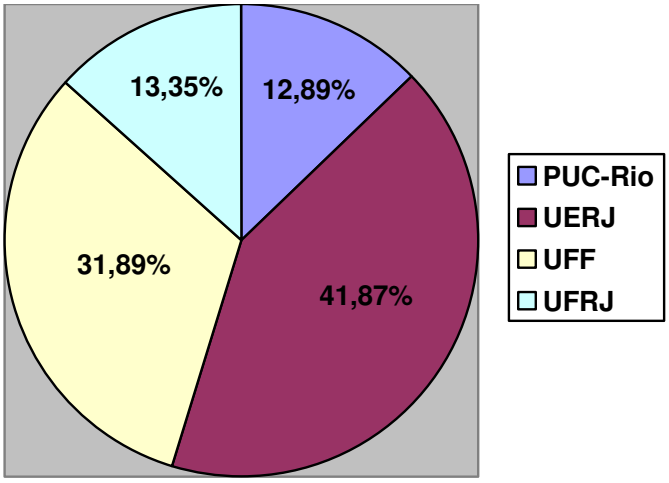

Figura 15: Distribuição do número de palavras no corpus por universidade

Ressalta-se novamente nesta seção a maior concentração de dados provenientes da UERJ (41,87%) e da UFF (31,89%). A PUC-Rio e a UFRJ apresentam quantidades menores. O que deve ser notado aqui, no entanto, é que apesar de haver um número menor de redações da UFRJ (12) do que da PUC-Rio (15) no corpus de pesquisa, conforme indicado na Seção 5.1.1.2, o primeiro grupo contém um número percentual maior de palavras.

Considerando que há 96 redações ao todo no corpus, conclui-se que cada redação contém em média 536 palavras. Pode-se, porém, verificar o número médio de palavras por redação por centro universitário como indicado na Figura 16.

PUC-Rio - Certificação Digital Nº 0610483/CA

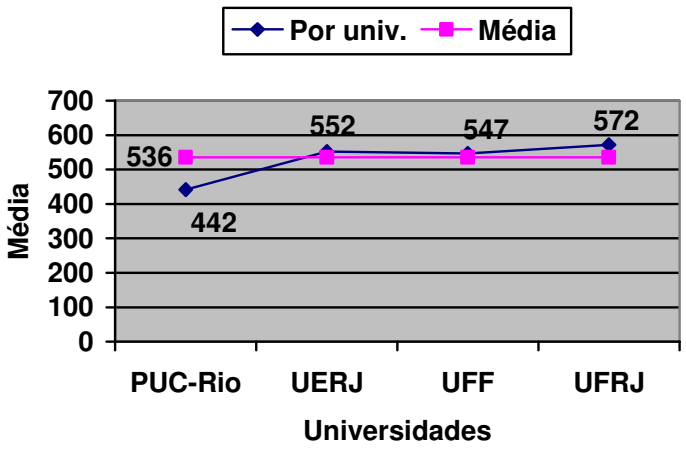

Figura 16: Média de palavras por redação em cada universidade

As redações escritas por alunos da UFRJ são as mais expressivas em termos de palavras, com uma média de 572. Ainda acima da média, encontram-se as redações escritas por alunos da UERJ (552 palavras) e os da UFF (547 palavras). As redações da PUC-Rio de uma forma geral se caracterizam por não conterem o número mínimo de palavras quando a média é calculada, ou seja, cada texto tem aproximadamente 442 palavras. Não se afirma, contudo, que todas as redações provenientes das outras três universidades contêm, no mínimo, 500 palavras. Verificou-se que algumas redações ficaram aquém do mínimo estabelecido enquanto outras passaram deste mínimo, mas nunca excedendo o limite máximo de 1.000 palavras. Não foram descartadas as redações que continham menos de 500 palavras, mas assegura-se que todas contêm, no mínimo, 300 palavras.

Em termos de tópicos escolhidos para a realização das redações, verifica-se que as mesmas são a respeito de 12 das 14 sugestões fornecidas. Não há, desta forma, nenhuma redação acerca do serviço militar (tópico G) ou do status atual da Europa (tópico M). A Figura 17 ilustra a distribuição das redações por tópico no corpus de pesquisa.

PUC-Rio - Certificação Digital Nº 0610483/CA

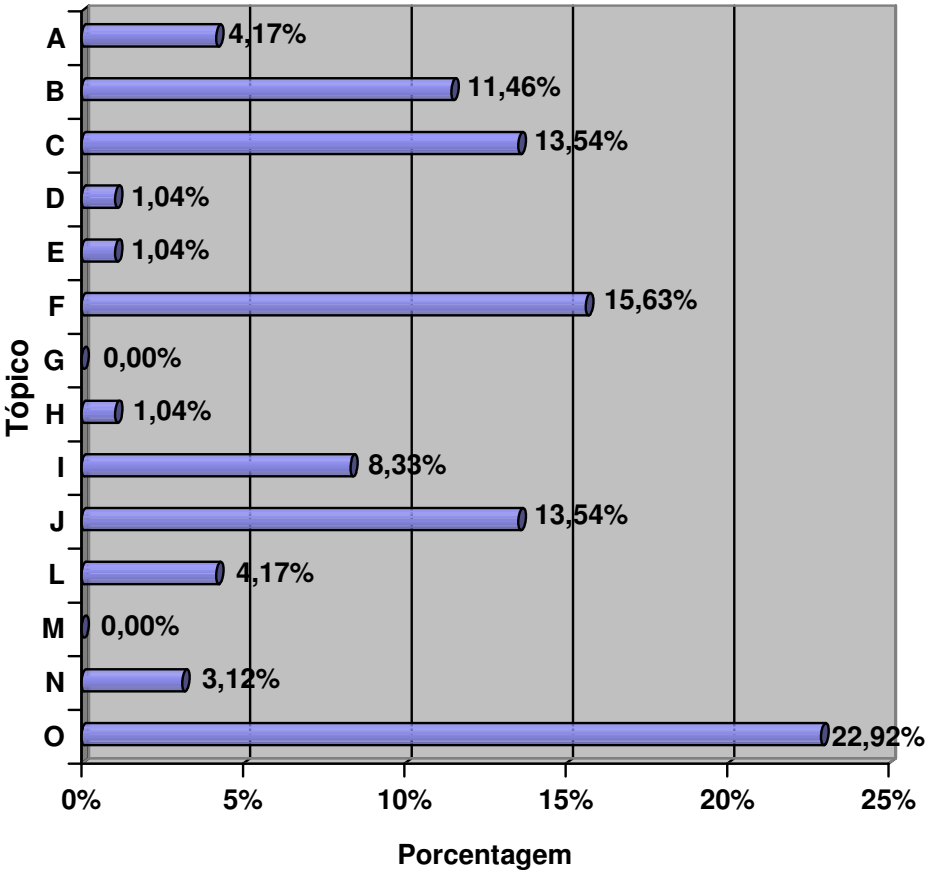

Figura 17: Distribuição de redações por tópicos

O tema mais freqüente é o que discute se há espaço para o sonho e a imaginação no mundo moderno (tópico O), totalizando 22,92% das redações. Em segundo lugar, encontram-se 15,63% dos textos acerca do status da televisão no final do século XX (tópico F). Com 13,54%, podem ser identificados o tópico C a respeito do valor de um curso universitário e o tópico J que discorre a respeito de igualdade na sociedade a partir de uma citação da obra *A Revolta dos Animais* de George Orwell. O tópico B – acerca do papel do sistema penitenciário – também parece ter sido um dos mais populares com 11,46%. Os outros nove tópicos aparecem com percentuais inferiores a 10%.

As redações também podem ser agrupadas em termos da época na qual a coleta foi realizada. Adotando-se o semestre como unidade temporal, têm-se cinco períodos de tempo que correspondem desde o ano de 2005 até a primeira metade de 2007. A distribuição de redações por tempo encontra-se representada na Figura 18.

PUC-Rio - Certificação Digital Nº 0610483/CA

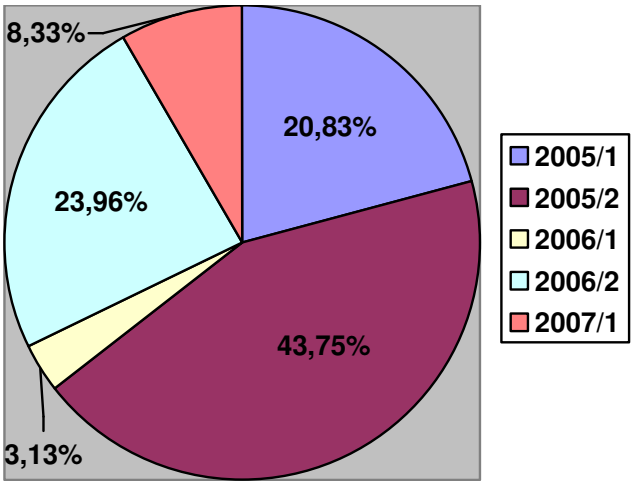

Figura 18: Distribuição de redações no Br-ICLE por época de coleta

PUC-Rio - Certificação Digital Nº 0610483/CA

A maior parte das redações (43,75%) foi coletada no segundo semestre de 2005. Com um pouco mais de um quinto das redações, tem-se o segundo semestre de 2006 (23,96%) e o primeiro de 2005 (20,83%). Por último, encontram-se os primeiros semestres de 2007 (8,33%) e 2006 (3,13%).

Por fim, deve-se caracterizar o corpus Br-ICLE de uma forma mais geral. Trata-se de um corpus de LE já que os participantes da pesquisa, autores das redações, são falantes brasileiros de inglês. Tem-se um corpus escrito monolíngüe já que o mesmo compreende redações redigidas em uma única língua, a saber, o inglês. Por conter um tipo textual bastante específico, ele também pode ser descrito como um corpus especializado. É um corpus sincrônico e contemporâneo visto que abrange um único período de tempo, qual seja, o atual. Como esta pesquisa tem como objetivo descrever o mesmo, o corpus pode ser descrito como de estudo. Em termos de tamanho, ele configura-se como pequeno nos termos de Berber Sardinha (2002), como explicado na Seção 2.1.1.

5.1.2. LOCNESS

O corpus de referência empregado nesta dissertação corresponde a um recorte do *Louvain Corpus of Native English Essays*, também conhecido por LOCNESS. Este corpus foi compilado pelo *Centre for English Corpus Linguistics* (CECL) da Universidade Católica de Louvain na Bélgica de forma a

permitir a comparação entre os diversos subcorpora nacionais do ICLE (cf. Seção 5.1.1) e a produção de falantes de IL1. Recski (2004, p. 5) e Baker, Hardie & McEnery (2006, p. 93) chegam a afirmar que o LOCNESS é um subcorpus do ICLE, o que parece ser inapropriado haja vista que o mesmo é de natureza distinta, não englobando redações de falantes de ILE.

O LOCNESS encontra-se disponível comercialmente para fins de pesquisa. Para tanto, deve-se efetuar o pagamento de 100 euros e o envio de uma ficha cadastral na qual são registrados os dados pessoais do comprador. Nesta ficha, também são indicadas informações acerca da pesquisa a ser realizada (título e tipo). A licença comprada é restritiva, sendo válida somente para a pessoa que adquire o corpus e para aqueles que trabalham diretamente com o comprador. Solicita-se também que se envie uma cópia de todo o material publicado a partir do corpus para o centro de pesquisa em questão e que o mesmo seja mencionado nas publicações.

O material de referência que acompanha o corpus indica que o mesmo contém 288.177 palavras. Os textos que integram o LOCNESS encontram-se subdivididos em quatro categorias:

(a) redações argumentativas escritas por alunos universitários americanos (149.574 palavras);

(b) redações acerca de temas literários produzidas por alunos universitários americanos (18.826 palavras);

(c) redações de cunho argumentativo ou literário escritas por alunos universitários britânicos (59.568 palavras);

(d) redações argumentativas redigidas por estudantes britânicos do *A-level*[94] (60.209 palavras).[95]

Com relação aos códigos empregados, indica-se que foram utilizadas três possibilidades sempre entre os sinais matemáticos de menor (<) e maior (>), sendo ignorados automaticamente pelo programa de análise textual aqui empregado (ver Seção 6.1 para mais detalhes a respeito do mesmo). Citações

[94] O *A-level* é um exame ao qual o aluno britânico se submete ao final do 13º ano de estudo e ainda antes de entrar na universidade.

[95] Uma descrição diferente do corpus pode ser encontrada no site de divulgação do mesmo (http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/locness1.htm) no qual se afirma haver 324.304 palavras no LOCNESS. A diferença de descrições, no entanto, não será discutida

PUC-Rio - Certificação Digital Nº 0610483/CA

longas e/ou não-integradas em frases foram substituídas por <*> para que a escrita desta população não fosse misturada com a de outro grupo distinto. Referências parentéticas a fontes consultadas foram igualmente removidas, mas utilizou-se o código <R> para marcar tal exclusão. Finalmente, a etiqueta <?> foi empregada para marcar trechos ilegíveis nas redações originais.

A escolha deste corpus justifica-se por razões práticas e teóricas. Em relação à praticidade da realização de um estudo contrastivo como o presente, coloca-se a impossibilidade de coleta de dados a distância em países nos quais o inglês é a L1. Neste sentido, a disponibilização comercial do LOCNESS revelou-se conveniente. Aliada a estas considerações, há uma justificativa de fundo teórico para o emprego deste corpus: o fato de haver vários estudos contrastivos entre os corpora do ICLE e o LOCNESS (Aarts & Granger, 1998; Altenberg & Tapper, 1998; Granger, 1998c; Granger & Rayson, 1998; Lorenz, 1998; Ringbom, 1998; Virtanen, 1998; Recski, 2004; entre outros). Assim sendo, tal escolha encontra-se respaldada na literatura específica apesar de o LOCNESS não conter todas as informações acerca das coletas realizadas e dos participantes de pesquisa como será indicado nas próximas seções.

PUC-Rio - Certificação Digital Nº 0610483/CA

5.1.2.1. Recorte do corpus

De forma a permitir a comparação entre os corpora, utiliza-se nesta dissertação um recorte do LOCNESS. Inicialmente, foram desconsideradas as redações produzidas por outros alunos que não universitários já que o foco da análise recai neste tipo de escrita. Desta forma, as redações do *A-level* foram excluídas da seleção final.

Como a composição do Br-ICLE contempla apenas redações argumentativas, optou-se por seguir este mesmo critério no LOCNESS. Portanto, as redações de cunho literário e/ou histórico-expositivas produzidas por alunos americanos e britânicos foram excluídas.

---

neste trabalho. O que se utilizou nesta dissertação foi um recorte deste corpus, que será explicado na Seção 5.1.2.1.

O tamanho (em número de total de palavras) também foi um critério considerado assim como no Br-ICLE (cf. Seção 5.1.1.5). As redações que continham menos de 300 palavras no LOCNESS foram desconsideradas.

De forma resumida, pode-se afirmar que com esta seleção buscou-se garantir, com maior confiabilidade, a comparabilidade entre os corpora.

5.1.2.2. Princípios de coleta de dados

Semelhante ao Br-ICLE, as redações que compõem o LOCNESS são de duas naturezas: argumentativas e literárias. No entanto, as últimas não entraram na seleção final deste estudo de forma a manter a comparabilidade entre os corpora conforme explicitado na Seção 5.1.2.1.

A língua na qual as redações foram escritas é igualmente a inglesa. Contudo, no LOCNESS, o inglês é a primeira língua dos participantes de pesquisa (cf. a descrição que acompanha o corpus), que também são estudantes de universidades americanas e britânicas.[96]

O tamanho da redação é semelhante ao do Br-ICLE, ou seja, geralmente entre 500 e 1.000 palavras, tendo as redações em média 821 palavras. As mesmas foram realizadas em condições distintas: seja com tempo determinado ou livre; parte integrante de uma prova ou como atividade extra; possibilidade ou proibição do uso de materiais de referência.[97]

Em relação ao comando dado para a realização das redações, não há nenhuma especificação no material descritivo que acompanha o LOCNESS. No entanto, os tópicos foram depurados a partir da leitura do corpus pelo pesquisador. Diferentemente do Br-ICLE, as redações do LOCNESS não são baseadas em temas tão específicos, havendo uma ampla gama destes. Somente quatro temas (A, D, I e L – cf. Seção 5.1.1.1) são repetidos no LOCNESS quando da coleta

[96] Não há informação a respeito do curso de graduação dos colaboradores; porém, dada a especificidade de alguns temas, especialmente os literários, supõe-se tratar de graduação em línguas e/ou literaturas. Além disto, não há registro do ano que estes alunos cursavam quando da contribuição para a pesquisa. No entanto, dado que o corpus foi compilado para fins de contraste com o ICLE, acredita-se que a população tenha características semelhantes.

[97] Dada a incompletude das informações no material que acompanha o corpus, acredita-se que os participantes não tenham preenchido um questionário semelhante ao empregado no ICLE. Isto impossibilita o conhecimento de inúmeras informações a respeito dos participantes e da coleta de dados em si. Em alguns momentos, as informações são apresentadas com pontos de interrogação, o que provavelmente indica a incerteza na parte descritiva.

PUC-Rio - Certificação Digital Nº 0610483/CA

realizada na *Indiana University*. Estes se relacionam a assuntos tais como o crime como uma atividade que não compensa, a relação entre percepção financeira e contribuição para a sociedade, o lado negativo da causa feminista, o dinheiro como a origem de males. Outros oito tópicos originalmente empregados no Br-ICLE guardam alguma relação com temas utilizados no LOCNESS:

(a) o sistema prisional (tópico B) é, de certa forma, retomado em redações a respeito da pena de morte;

(b) a questão do ensino universitário (tópico C) encontra espaço na discussão de se professores não-falantes de IL1 devem ensinar alunos que são falantes de IL1;

(c) o papel da censura (tópico E) também está presente em discussões acerca da liberdade de imprensa e a censura da mídia televisiva;

(d) a importância da televisão (tópico F) é debatida em algumas redações acerca das grandes invenções do século XX;

(e) a possível diferenciação entre seres humanos (tópico J) é retomada em redações sobre o racismo;

(f) o debate acerca da unificação européia (tópico M) é encontrado em redações do LOCNESS, mas estas enfocam a Grã-Bretanha;

(g) a natureza (tópico N) é revisitada em tópicos como a poluição aquática;

(h) a questão do mundo moderno, dominado por ciência, tecnologia e industrialização (tópico O) pode ser associada ao tema das grandes invenções do último século e seus impactos para a população em geral.

Apesar de ser possível notar algumas semelhanças entre os tópicos, isto não implica a afirmação de que os mesmos são iguais em ambos os corpora. A título de exemplificação, citam-se alguns outros temas de redações do LOCNESS: (a) aborto, (b) ação afirmativa, (c) AIDS, (d) controle de armas, (e) divórcio, (f) energia nuclear, (g) estereotipagem de cores, (h) ética, (i) eutanásia, (j) ioga, (l) legalização da maconha, (m) reciclagem, (n) religião nas escolas e (o) veiculação de conteúdo sexual e/ou violento na mídia. Estes temas não encontram nenhum correspondente no Br-ICLE.

PUC-Rio - Certificação Digital Nº 0610483/CA

5.1.2.3.
Coletas de dados

As redações foram coletadas em dois países distintos, a saber, Grã-Bretanha e Estados Unidos. Ao todo, foram envolvidas, no mínimo, seis universidades.

No caso britânico, não há indicação em que universidade(s) a(s) coleta(s) foi(foram) feita(s). Há, nesta parte do corpus, 33 redações que não integraram provas. Estas versavam a respeito da soberania da Grã-Bretanha em face à unificação da Europa.

Nos Estados Unidos, a coleta foi realizada em cinco instituições de ensino superior distintas: *Indiana University* (IND), *University of Michigan* (MIC), *Marquette University* (MRQ), *Presbyterian College* (PRB) e *University of South Carolina* (SCU). A Figura 19 indica o número de participantes por universidade.

PUC-Rio - Certificação Digital Nº 0610483/CA

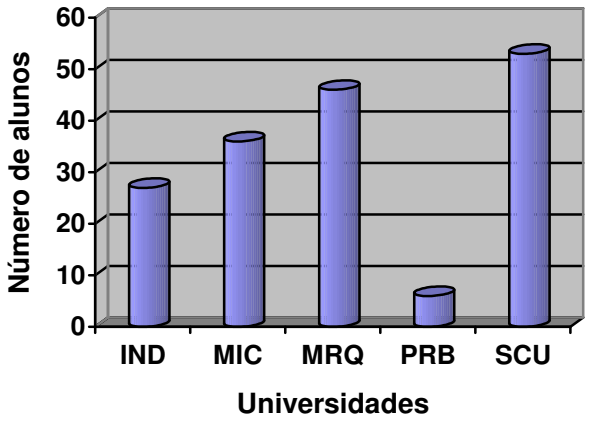

Figura 19: Distribuição de alunos por universidade (LOCNESS / Dados americanos)

As instituições SCU (53) e MRQ (46) foram as que concentraram o maior número de colaboradores. Juntas elas representam 58,92% das redações americanas coletadas. As universidades MIC e IND totalizaram, respectivamente, 36 e 27 participantes. Na outra extremidade, está a PRB com apenas 6 participantes.

5.1.2.4.
Condições de realização das redações

No caso britânico, as redações não foram realizadas sob controle temporal rígido. O que se registra neste caso é que as mesmas foram completadas em torno de uma hora. Nenhum material de referência foi empregado nesta coleta.

Na MIC, as redações foram realizadas com tempo marcado, não tendo havido consultas a materiais de referência.

As condições de realização dos textos na MRQ e na PRB foram opostas já que não houve delimitação de tempo e materiais extras puderam ser consultados. Entre estes recursos, são indicados livros e periódicos.

No caso da SCU, não é possível encontrar um padrão haja vista que as quatro coletas parecem ter sido feitas de formas diferentes e nem todas as condições foram registradas.

Situação semelhante concerne à IND na qual não há referência precisa se a tarefa de redação teve seu tempo limitado e se materiais de referência foram consultados.

PUC-Rio - Certificação Digital Nº 0610483/CA

5.1.2.5. Formatação do corpus

O LOCNESS é comercializado em formato digital – mais especificamente, TXT. Desta forma, não foi preciso digitar as redações coletadas. No entanto, uma desvantagem está intimamente ligada a esta economia de tempo: não é possível verificar nos arquivos originais ou nas redações em papel o que os participantes efetivamente escreveram.

O emprego de etiquetas no LOCNESS é igual ao Br-ICLE (cf. Seção 5.1.1.4). As marcações utilizadas servem para omitir informações desnecessárias ou para indicar a impossibilidade de compreensão do que foi escrito (ver Seção 5.1.2 para mais detalhes).

O fato de o LOCNESS já ser disponibilizado digitalmente não indica que o mesmo está pronto para o uso. As redações vieram em 14 arquivos TXT. Houve um trabalho inicial de separação de todas as redações de forma a manter um texto por arquivo, visando à posterior identificação da origem da linha de concordância quando do procedimento analítico (ver Seção 6.1.3).

Ao realizar a tarefa descrita anteriormente, percebeu-se que muitos arquivos não haviam sido processados adequadamente. Os códigos empregados para citações e referências não haviam sido inseridos em vários casos. A marcação de parágrafos também não era clara o suficiente em algumas redações, com quebras de parágrafo ao final de cada linha. Em uma das redações, havia uma parte final

com os dados do participante, tais como nome, idade, língua materna, curso universitário e respectivo período.

A nomeação dos arquivos no LOCNESS não foi tão minuciosa quanto a do Br-ICLE pelo fato de muitos dados não estarem disponíveis, como comentado anteriormente. Optou-se, então, por seguir, de certa forma, a nomeação dada pelos compiladores do mesmo. Os nomes dos arquivos compõem-se, então, de quatro partes:

(a) a letra B ou U, indicando o país de origem da redação: Grã-Bretanha ou Estados Unidos;

(b) uma seqüência de três letras, referindo-se à universidade: IND, MIC, MRQ, PRB, SCU e SUR[98];

(c) um número de dois dígitos que identifica a redação;

(d) um traço seguido de um número de um dígito que indica o número da coleta realizada.

A título de exemplificação, o código USCU09-3 indica que a mesma foi escrita por um aluno norte-americano da *University of South Carolina*. Esta é a nona redação da terceira coleta realizada neste centro universitário.

5.1.2.6. Caracterização do corpus

O recorte do LOCNESS empregado nesta dissertação totaliza 165.135 palavras e 11.146 palavras distintas em 201 textos (cf. Anexo 10.9). Como o corpus contempla dados de dois grupos nacionais distintos (cf. Seção 5.1.2), pode-se verificar a contribuição de cada um deles por número de palavras.

[98] No caso da Grã-Bretanha, onde não há informação acerca da universidade, os compiladores utilizaram esta seqüência para identificá-las.

PUC-Rio - Certificação Digital Nº 0610483/CA

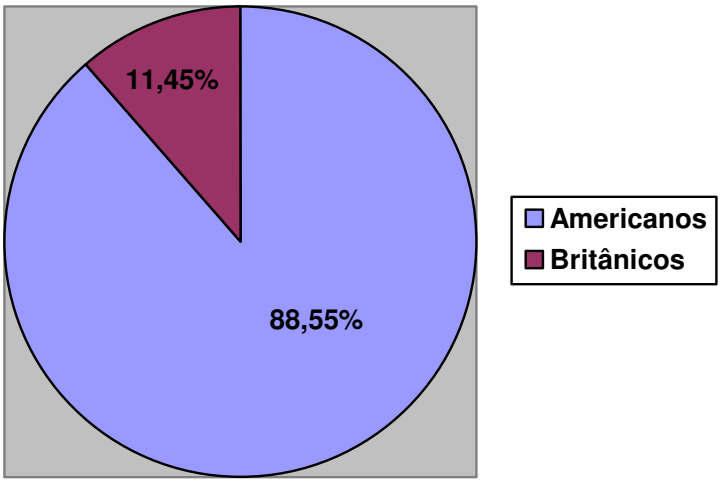

Figura 20: Distribuição do número de palavras no corpus por nacionalidade

Como ilustrado na Figura 20, os alunos americanos são responsáveis pela maior fatia do corpus: 88,55%. Os britânicos contribuem com uma menor parcela de redações: apenas 11,45%. Em termos de número de redações, há 168 produzidas por americanos e apenas 33 por britânicos (cf. Anexos 10.10 e 10.11).

Verifica-se também que a média de palavras por redação no LOCNESS não é a mesma para cada grupo nacional como indica a Figura 21.

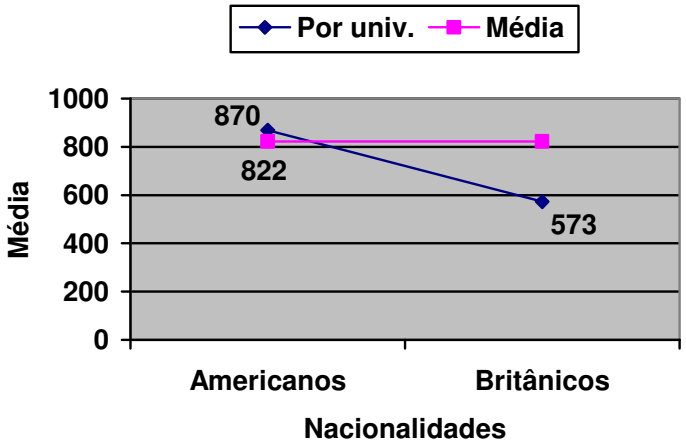

Figura 21: Média de palavras por redação por nacionalidade

Além de contribuir com um número menor de redações, as redações dos participantes britânicos também contêm um número reduzido de palavras. Em média, eles utilizam 573 palavras em suas redações ao passo que os participantes

PUC-Rio - Certificação Digital Nº 0610483/CA

americanos empregam 870. A média geral do corpus é de 822 palavras por redação.

Em relação aos tópicos, não é possível traçar um perfil completo de distribuição das redações pela razão discutida na Seção 5.1.2.2. Contudo, nota-se que os dois temas mais recorrentes são os que versam a respeito de grandes invenções e seus impactos na sociedade contemporânea (nos dados americanos), e a perda de soberania da Grã-Bretanha em face à unificação européia (nos dados britânicos), com 36 e 33 redações respectivamente. Outros tópicos menos freqüentes, mas recorrentes lidam com (a) crimes como atos não recompensadores (10 redações); (b) dinheiro como causa de males (7 redações); e (c) papéis de gêneros na sociedade (7 redações).

Em relação à época na qual a coleta de dados foi realizada, percebe-se a existência de quatro grupos semestrais, representados na Figura 22.

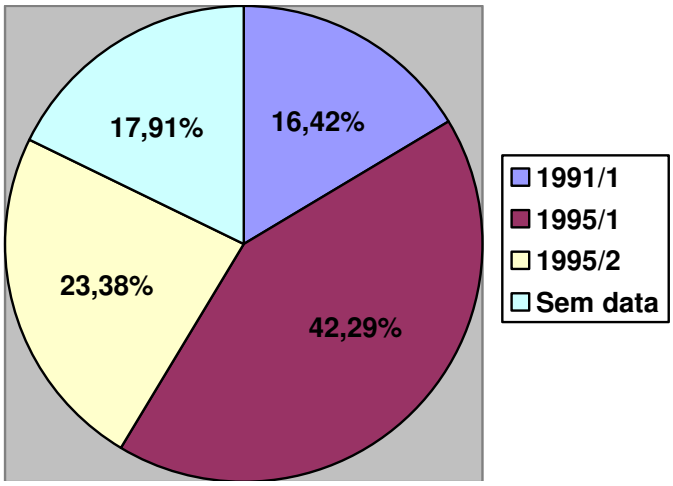

Figura 22: Distribuição de redações no LOCNESS por época de coleta

A maior parte da coleta foi realizada no primeiro semestre de 1995, totalizando 42,29% dos dados. O segundo maior grupo é o relativo ao segundo semestre deste mesmo ano com 23,38% das redações. Isto resulta que a maior parte da coleta (65,67%) foi realizada neste ano, com uma pequena parte coletada em 1991 (16,42%) e outra parcela sem data especificada (17,91%).

Por fim, pode-se caracterizar de forma mais geral o LOCNESS como um corpus de L1, ou seja, que contempla dados produzidos por falantes de inglês como primeira língua. Ele é um corpus escrito monolíngüe dado que o mesmo

PUC-Rio - Certificação Digital Nº 0610483/CA

contém somente redações em uma única língua. Pode ser caracterizado com um corpus especializado (Granger & Tribble, 1998, p. 204), que reflete a escrita de alunos universitários americanos e britânicos. Em relação ao tempo, ele é sincrônico e contemporâneo, tendo sido coletado entre os anos de 1991 e 1995. Como o seu uso neste estudo é o de prover parâmetros para comparação com o Br-ICLE, ele é descrito como de referência. Em relação ao seu tamanho, ele é descrito como pequeno-médio (cf. classificação proposta por Berber Sardinha, 2002 explicitada na Seção 2.1.1) já que ultrapassa 80.000 palavras.

5.2. Participantes

Além da descrição dos corpora, faz-se necessário também descrever os participantes da pesquisa, que contribuíram com as redações. Ao conhecer este perfil, pode-se ter uma noção mais geral de quem os corpora são representativos.

5.2.1. Br-ICLE

Como explicado na Seção 5.1.1.1, os participantes da pesquisa foram solicitados a preencher um questionário no qual deveriam registrar algumas informações pessoais. Esta seção é baseada na compilação das respostas registradas no questionário utilizado na pesquisa, que se encontra reproduzido no Anexo 10.3.

Em relação à nacionalidade, há somente uma resposta: todos os 96 participantes são brasileiros. Em termos de sexo, há uma maior concentração de participantes do sexo feminino (82,29%) em comparação aos participantes do sexo masculino (17,71%). A idade dos participantes varia em faixas que vão desde até 18 até 50 anos como indica a Figura 23.

PUC-Rio - Certificação Digital Nº 0610483/CA

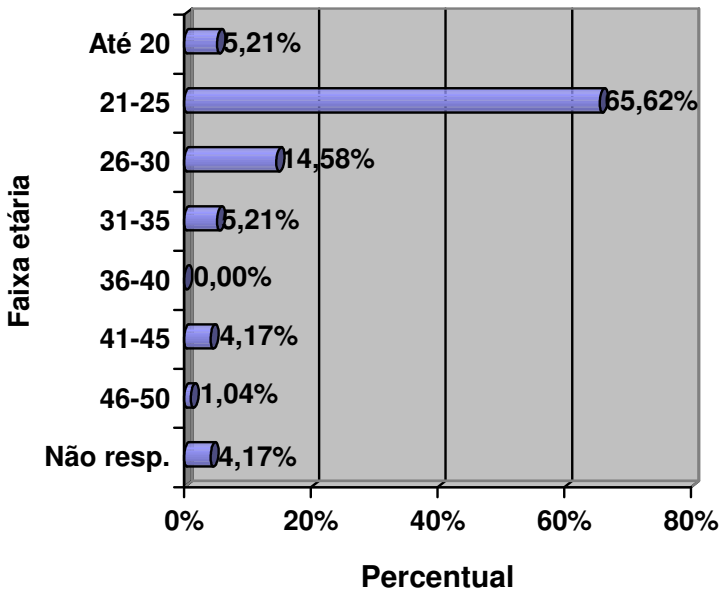

Figura 23: Distribuição dos participantes por faixa etária

A grande maioria dos alunos (65,62%) tem entre 21 e 25 anos. A segunda maior concentração de alunos (apenas 14,58%) encontra-se na faixa etária seguinte de 26 a 30 anos. Ressalta-se que 4,17% dos participantes não responderam a este item do questionário. De uma forma geral, a média de idade é de 25 anos, o que difere da média do ICLE como um todo, que é de 20 anos (Granger, 1998c, p. 9).

Quando a língua materna é considerada, nota-se que os alunos são falantes de português como primeira língua em sua grande maioria (98,96%). Apenas um aluno declarou ter tanto o português como o alemão como suas línguas maternas. Os pais dos participantes também têm o português como língua materna (96,88%). Ao se analisar o perfil das mães destes alunos, percebe-se mais uma vez que o português é a principal língua materna (95,84%).

Em casa, estes alunos utilizam somente a língua portuguesa para se comunicar em grande parte das vezes (91,67%). Há oito casos nos quais os alunos empregam línguas diversas para se expressar em casa; porém, em todos eles, o português está presente em igual ou maior proporção.

Quando questionados a respeito do possível comando de outras línguas estrangeiras diferentes do inglês, aproximadamente metade dos alunos respondeu

PUC-Rio - Certificação Digital Nº 0610483/CA

que não tinha conhecimento de nenhuma outra língua como se pode notar na Figura 24.

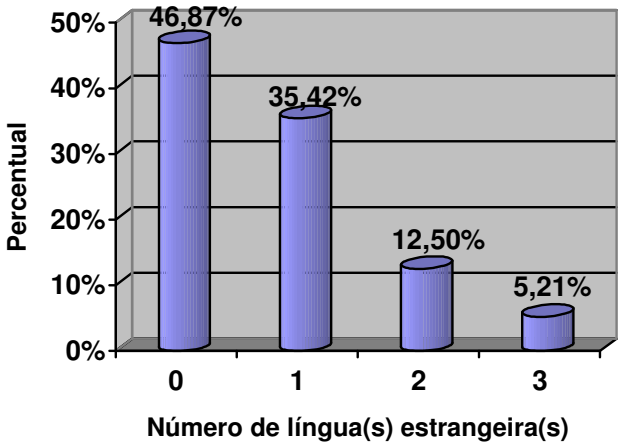

Figura 24: Conhecimento de línguas estrangeiras diferente do inglês

A outra metade dos alunos dividiu-se entre o conhecimento de uma (35,42%), duas (12,50%) ou três (5,21%) línguas estrangeiras. Contudo, o valor percentual decresce à medida que o número de línguas aumenta.

Investigou-se se os alunos foram instruídos formalmente em língua inglesa durante os ensinos fundamental e médio. Nota-se que o ensino fundamental foi realizado majoritariamente em português (96,88%). Apenas três alunos estudaram em escolas nas quais o ensino era realizado em outras línguas. Mesmo assim, em somente um caso, o aluno não teve contato com a língua portuguesa neste segmento de ensino. No caso do ensino médio, há uma nova preponderância da língua portuguesa (95,84%) à semelhança do que foi descrito para o ensino fundamental. Mais uma vez ressalta-se que um número reduzido (somente dois alunos) não teve contato com a língua portuguesa durante esta fase do ensino escolar. Em termos de exposição formal à língua inglesa no contexto escolar (aqui entendido como qualquer um dos segmentos), os alunos declararam ter tido em média 6,94 anos de ensino desta língua.

Alguns aspectos acerca do ensino universitário destes alunos também foram registrados no questionário. Comenta-se inicialmente o curso de origem destes alunos: (a) inglês e respectivas literaturas, (b) português, inglês e respectivas

PUC-Rio - Certificação Digital Nº 0610483/CA

literaturas e (c) tradução (português-inglês). A Tabela 1 indica a distribuição dos alunos nestes três cursos.

Tabela 1: Distribuição de alunos por curso de graduação

Curso	Percentual
Português, inglês e respectivas literaturas	56,25%
Inglês e respectivas literaturas	40,63%
Tradução (português-inglês)	3,12%

A Tabela 1 deve ser, no entanto, interpretada à luz da Figura 11 presente na Seção 5.1.1.2, que se encontra reproduzida a seguir em forma de tabela.

Tabela 2: Distribuição de alunos por universidade (Br-ICLE)

Universidade	Percentual
PUC-Rio	15,62%
UERJ	40,63%
UFF	31,25%
UFRJ	12,50%

A grande concentração de alunos de alunos do curso de habilitação dupla em português e inglês (56,25%) deve-se ao fato de tal grupo reunir alunos da PUC-Rio, UFF e UFRJ. Mais especificamente, as duas últimas universidades só oferecem este tipo de curso. Já a concentração de alunos do curso de habilitação única em língua inglesa deve-se ao fato de o maior grupo colaborador da presente pesquisa (39 ao todo) ser proveniente da UERJ, universidade na qual somente esta modalidade é oferecida em seu campus principal. O curso de tradução, que soma apenas 3,12%, representa somente uma fatia menor da coleta realizada na PUC-Rio, única das quatro universidades a oferecer tal possibilidade.

Em relação ao ano em curso, nota-se que os participantes concentram-se na segunda metade de seus cursos de origem, a saber, terceiro, quarto ou quinto anos conforme indicado na Tabela 3.

Tabela 3: Quantidade de anos cursados na universidade

Anos	Percentual
Três	31,25%
Quatro	38,54%
Cinco	30,21%

PUC-Rio - Certificação Digital Nº 0610483/CA

Há uma concentração de alunos no quarto ano de seus respectivos cursos (38,54%); porém, esta não difere muito dos outros grupos que se encontram no terceiro e quinto anos que totalizam 31,25% e 30,21%, respectivamente. Na verdade, a concentração dos alunos nestes três anos decorre de um dos princípios para a coleta de dados, que exige que os alunos estejam no terceiro ou quarto anos de seus cursos de forma a considerá-los estudantes avançados de inglês (cf. Seção 5.1.1.1).

Todos os participantes tinham aulas em duas línguas distintas, a saber, inglês e português. No entanto, a população investigada diferencia-se quanto ao número de anos de instrução formal de inglês na universidade como indica a Figura 25.

PUC-Rio - Certificação Digital Nº 0610483/CA

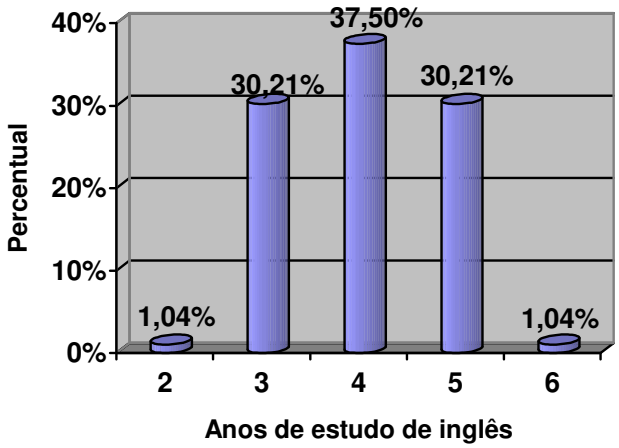

Figura 25: Anos de estudo de inglês na universidade

Grande parte dos alunos (37,50%) teve quatro anos de estudo de inglês na universidade quando da escritura da redação. Há, também parcelas consideráveis de alunos que tiveram 3 ou 5 anos de instrução universitária formal de língua inglesa, que totalizam 30,21% cada. Há somente dois casos destoantes nos quais os alunos indicam terem tido 2 e 6 anos de ensino de inglês.

Também foram registradas as possíveis viagens realizadas pelos participantes da pesquisa a países cuja língua oficial é o inglês. Dentre todos os participantes, 39,58% afirmaram terem viajado para tais países ao passo que 60,42% indicaram não ter tido tal experiência.

5.2.2.
LOCNESS

Esta seção enfoca a descrição dos alunos envolvidos na coleta de dados para o LOCNESS. Como não há acesso aos questionários de perfil (nem se sabe se os mesmos foram utilizados), o conhecimento deste grupo é mais reduzido do que o do Br-ICLE e também não é possível descrever padrões gerais em muitos dos casos.

Como indicado na Seção 5.1.2.3, os participantes são alunos universitários de duas nacionalidades diferentes (americanos e britânicos) e provenientes de, no mínimo, seis instituições de ensino superior distintas (IND, MIC, MRQ, PRB, SCU e SUR), sendo cinco dos Estados Unidos e uma da Grã-Bretanha.[99]

Com relação à idade dos alunos, parece haver uma convergência nas coletas realizadas:

(a) MIC – variação entre 19 e 23 anos;

(b) MRQ – variação entre 18 e 21 anos (exceções: três participantes com 30, 31 e 40 anos);

(c) PRB – variação entre 20 e 22 anos;

(d.1) USC – variação entre 17 e 20 anos (exceções: dois participantes com 34 e 66 anos);

(d.2) USC – variação entre 18 e 19 anos (exceção: um participante com 28 anos);

(d.3) USC – variação entre 17 e 19 anos (exceção: um participante com idade não-declarada);

(d.4) USC – variação entre 17 e 21 anos (exceções: três participantes como 39, 41 e 57 anos).

De forma resumida, tem-se que a idade da maioria destes alunos varia entre 17 e 23 anos. No entanto, a coleta realizada na IND, com 27 participantes, difere das anteriores uma vez que a variação ocorre entre 22 e 48 anos.

Em relação à língua inglesa, todos os alunos são falantes de IL1, utilizando esta língua na comunicação realizada em casa e tendo pais e mães que também são falantes de IL1, havendo apenas quatro exceções.

PUC-Rio - Certificação Digital Nº 0610483/CA

5.3.
Resumo

Esta seção ofereceu uma minuciosa descrição do corpus de pesquisa – o Br-ICLE – tanto a respeito da compilação do corpus propriamente dito quanto dos participantes da pesquisa. Esta última parte adveio da análise dos questionários de perfil preenchidos pelos alunos que gentilmente aceitaram colaborar com a pesquisa. Procedimento semelhante foi adotado para o LOCNESS, o corpus de referência, que contempla dados produzidos por alunos americanos e britânicos.

Ressalta-se, no entanto, a diferença na descrição dos dois corpora. O de pesquisa, foco da análise aqui realizada é mais detidamente analisado uma vez que todos os dados encontram-se disponíveis. Neste caso, o pesquisador participou tanto da coleta de grande parte dos dados e do tratamento de todas as redações aqui investigadas. Em relação ao de referência, tendo em vista a impossibilidade de se compilar o mesmo a distância, optou-se por se adquirir um corpus que pudesse ser comparado ao mesmo.

Apesar de várias pesquisas terem empregado o LOCNESS como um corpus comparável aos diferentes subcorpora do ICLE (ver Granger, 1998a, por exemplo), houve a necessidade de se recorrer a um recorte do mesmo nesta dissertação. A exclusão de textos foi realizada com base em critérios definidos (redações não produzidas por universitários; de natureza literária e/ou histórico-expositiva; e/ou com menos de 300 palavras) de forma tal que a comparabilidade entre os corpora pudesse ser assegurada e que se reduzisse a possibilidade de outros fatores não-considerados nesta pesquisa influenciassem no resultado final.

O que se tem são dois corpora que contemplam redações escritas por universitários, estudantes avançados de inglês. A principal diferença entre as populações concerne ao status que as mesmas têm em relação à língua inglesa: para os colaboradores do Br-ICLE a mesma é uma língua estrangeira que eles estudam em nível superior enquanto para os colaboradores do LOCNESS ela representa sua língua materna. Há outras variáveis entre os sujeitos de pesquisa, tais como sexo, língua materna dos pais, conhecimento de outras línguas estrangeiras, que não serão consideradas nesta investigação.

[99] Não se pode ter certeza da descrição para o caso britânico visto que a informação não é fornecida na documentação que acompanha o LOCNESS. Por este motivo, toda a descrição

PUC-Rio - Certificação Digital Nº 0610483/CA

Por fim, o tamanho de ambos os corpora chama a atenção quando comparado a corpora de dimensões maiores, como o BNC que contém 100 milhões de palavras e o BoE que contempla 524 milhões de palavras. No entanto, a compilação de corpora tais como os utilizados aqui não é fácil tendo em vista a dificuldade de se conseguir os dados necessários. Assim sendo, como afirma Granger (1998c, p. 10), citada na epígrafe deste capítulo, corpora que contemplam textos de falantes de L2 / LE muito provavelmente não chegarão ao tamanho destes grandes corpora. Além disto, estes corpora são especializados, o que os difere de corpora gerais de tamanhos gigantescos.

PUC-Rio - Certificação Digital Nº 0610483/CA

contida nesta seção enfoca somente os dados dos participantes americanos.