GISELE MONTILHA PINHEIRO

***Redações do ENEM: estudo dos desvios da norma padrão
sob a perspectiva de corpus***

São Paulo
2007

GISELE MONTILHA PINHEIRO

**Redações do ENEM: estudo dos desvios da norma padrão
sob a perspectiva de corpus**

Dissertação apresentada à Faculdade de Filosofia, Letras
e Ciências Humanas da Universidade de São Paulo para
obtenção do título de Mestre.

Área de Concentração: Estudos Lingüísticos e Literários
em Inglês

Orientadora: Profa. Dra. Stella E. O. Tagnin

São Paulo
2007

P654

Pinheiro, Gisele Montilha

Redações do ENEM : estudo dos desvios da norma padrão sob a perspectiva de corpus / Gisele Montilha Pinheiro ; orientadora Stella E. O. Tagnin. -- São Paulo, 2007.
152 f.

Dissertação (Mestrado - Programa de Pós-Graduação em Estudos Lingüísticos e Literários em Inglês. Área de concentração: Estudos Lingüísticos e Literários em Inglês) - Departamento de Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

1. Lingüística aplicada. 2. Lingüística de corpus. 3. Português do Brasil. 4. Escrita (Avaliação). 5. Lingüística computacional. 6. Língua culta (Desvios). 7. Exame Nacional de Ensino Médio (2002) – Brasil. I. Título.

21ª. CDD 418

FOLHA DE APROVAÇÃO

Gisele Montilha Pinheiro

Redações do ENEM : estudo dos desvios da norma padrão sob a perspectiva de corpus

Dissertação apresentada à Faculdade de Filosofia,
Letras e Ciências Humanas da Universidade de São
Paulo para obtenção do título de Mestre

Área de Concentração: Estudos Lingüísticos e
Literários em Inglês

Aprovado em: _____

Banca Examinadora

Prof. Dr.: _____

Instituição: _____ Assinatura: _____

Prof. Dr.: _____

Instituição: _____ Assinatura: _____

Prof. Dr.: _____

Instituição: _____ Assinatura: _____

Prof. Dr.: _____

Instituição: _____ Assinatura: _____

Prof. Dr.: _____

Instituição: _____ Assinatura: _____

Prof. Dr.: _____

Instituição: _____ Assinatura: _____

Prof. Dr.: _____

Instituição: _____ Assinatura: _____

Prof. Dr.: _____

Instituição: _____ Assinatura: _____

DEDICATÓRIA

À MARIA PERPÉTUA DA CONCEIÇÃO, minha mãe, luz da minha vida sempre, à *WILMA CONCEIÇÃO MONTILHA*, querida irmã, pela confiança e dedicação insuperáveis, às minhas filhas *BETÂNIA* e *LUANA*, tanta espera, tanta paciência comigo... Este trabalho é por vocês.

AGRADECIMENTOS

À Profa. Dra. Stella E.O. Tagnin, pela confiança em meu trabalho e todo o apoio nesse período de orientação.

À Profa. Dra. Sandra M. Aluísio, por não me deixar desistir de um sonho e pelo auxílio inestimável no curso deste trabalho.

Ao NILC, por me introduzir no universo dos estudos lingüístico-computacionais; às suas equipes, de hoje e do passado, por fazerem realmente parte da minha vida.

À Diretoria Regional de Ensino de Osasco, na figura de sua dirigente, Maria de Fátima Carnellós.

A Deus e toda a fraternidade espiritual que me guia.

À família maravilhosa que tenho, em especial aos meus irmãos: Hélio, Silvia e Wilma.

RESUMO

PINHEIRO, G.M. **Redações do ENEM: estudo dos desvios da norma padrão sob a perspectiva de corpus**. 2007, 140 f. Dissertação – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2007.

Desvios da norma padrão, comumente chamados de “erros”, são fatos comuns na escrita dos aprendizes da variante culta de uma língua materna como o português brasileiro. Tratados como um “mal a ser combatido”, eles são, na verdade, importantes indícios do processo de assimilação da escrita culta pelo falante nativo. Revelam qual a tendência da transformação que naturalmente ocorre numa língua, demonstrando, por exemplo, a obsolescência das gramáticas tradicionais, que não aceitam determinadas construções já muito frequentes. Mas seria possível detectar algum padrão desses desvios? Haveria desvios típicos de um determinado perfil de redatores? Essas indagações motivaram a presente investigação, que se baseou na concepção de que esses estudos são de natureza empírica, comprometidos com a noção de que a língua funciona tal como um sistema probabilístico, de onde é possível prever *tendências*, por exemplo, de mudança. Falamos, pois, de uma investigação à luz da Linguística de Corpus. Composto de redações do Exame Nacional do Ensino Médio (ENEM), edição de 2002, cedidas pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) juntamente com determinados traços do perfil dos redatores, construímos um corpus que foi batizado de *Corvo*, e se ocupou de uma faixa específica de textos: a de pior desempenho no ENEM no quesito domínio da norma culta. Observamos, desse modo, textos em que, supostamente, há frequência maior de desvios e maior variedade de tipos de desvios. Nossa metodologia de pesquisa apoiou-se no uso do revisor gramatical automático *ReGra*, bastante popular no país e que auxilia o usuário no uso correto do português culto padrão. Além disso, construímos um material próprio de detecção e classificação dos desvios gramaticais, aumentando a capacidade de tratamento automático dos textos. Assim, foi possível gerar uma versão do corpus anotada em desvios, i.e., os textos apresentam indicações de quando e qual tipo de desvios ocorrem. Como resultado temos um mapeamento do *Corvo*; ou seja, um panorama dos desvios típicos de um determinado tipo de perfil de redator. Constatamos a deficiência ortográfica como o traço típico do grupo de indivíduos investigado, mas, sobretudo, que a ortografia é motor para o pleno funcionamento de uma revisão gramatical automática. O revisor *ReGra* mostrou-se incapaz de processar satisfatoriamente textos desse tipo de redator, mas, ainda assim, comprovou que esses textos apresentam desvios gramaticais de tratamento complexo, cuja intervenção do revisor, se acontece, pouco altera na qualidade geral dos mesmos. Com respeito à tipologia de desvio, pudemos constatar a validade da tipologia aplicada na pesquisa, que advém do *ReGra* e, portanto, está à margem das discussões teóricas ortodoxas. De fato, há recorrência de tipos de desvios, e isso numa frequência que nos autoriza admitir a fraca assimilação de certas regras gramaticais tomadas como básicas (p.ex., a pontuação, a concordância e a regência). Constatamos, com relação ao perfil de redatores, que textos com maior potencial para a revisão da escrita, i.e., aqueles que alteram significativamente a qualidade textual com interferências pontuais de revisão, são justamente os produzidos pelos concluintes do ensino médio e não pelos egressos.

Palavras-chave: Linguística de Corpus; redações do ENEM; desvios da norma-padrão; revisão gramatical automática; anotação de corpus.

ABSTRACT

PINHEIRO, G.M. **ENEM essays: a study of deviations from the standard norm from a corpus perspective**. 2007, 140 p. Masters' thesis – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2007.

Deviations from the standard norm, usually called 'mistakes', are common events in writing pieces of language learners speakers of a native language such as Brazilian Portuguese. They are treated as 'an evil that must be fought'. They are, in fact, important evidence of the acquisition process of writing in the standard norm by the native speaker. They reveal the transformation trend, which normally occurs in a language, showing, for instance, the obsolescence of traditional grammars that do not accept certain patterns, which are frequent nonetheless. However, is it possible to identify a pattern in these deviations? Are there common deviations among a certain profile of students? These are the questions motivating this study, which is based on the concept that these investigations are empirical in nature, and are marked by the notion that language operates as a probabilistic system, in which it is possible to forecast *trends* of change, for example. We are, therefore, speaking of an investigation in the light of Corpus Linguistics. We compiled a corpus of essays written during the National Middle Education Exam (ENEM) carried out in 2002. These essays were obtained with the National Institute for Research in Education (INEP) together with the profile of the students. The corpus was called *Corvo*, and it is made up of a certain bracket of texts: those having obtained the worst performance rate in the ENEM in the standard norm category. We observed, therefore, texts in which there are, allegedly, a greater frequency of deviations and a greater variety of kinds of deviations. The research methodology was supported by the electronic grammar checker – *ReGra* – which is very popular in the country, and helps the user in writing standard Portuguese correctly. In addition, we built a specific tool for identifying and classifying grammar deviations, thus, increasing the ability to treat the texts electronically. Therefore, it was possible to generate an annotated version of the corpus according to the deviations, i.e., the texts were annotated according to when and what kind of deviations they presented. As a result, we have a mapping of the *Corvo*; that is, a view of the common deviations of students belonging to a certain profile. We identified poor spelling as a common feature of the group, but, above all, that spelling is the engine enabling a full grammatical check to operate. The ReGra grammar checker was not able to satisfactorily process these kinds of texts, but, even so, it proved that these texts presented complex grammar deviations, and the intervention of the checker, when it is applied, little alters their overall quality. In regard to deviation typology, we identified the validity of the typology used in this study, which results from the ReGra and, therefore, lies in the outskirts of orthodox theoretical discussions. In fact, certain kinds of deviations reoccur at a frequency that enables us to admit a poor assimilation of certain grammatical rules considered basic (e.g., punctuation, agreement and use of prepositions). We found that in regard to the profile of the students, texts with a greater writing check potential, that is, a check that would significantly improve text quality through individual checking interferences, are texts produced by students who finished middle education and not those produced by students who are finishing the course.

Keywords: Corpus linguistics; ENEM essays; deviations from the standard norm; electronic grammar checker; corpus annotation

SUMÁRIO

I. Introdução.....	5
II. Desvios da norma padrão e revisão lingüística automática.....	8
II.1 Das noções de <i>norma padrão</i> e de <i>desvio</i>	8
II.1.1 Referências da norma padrão no Brasil e a “falência” do português culto	11
II.1.2 Desvios da norma padrão e a mensuração do desempenho do redator	14
II.2 Das noções de <i>desempenho</i> e de <i>revisão automática</i>	17
II.2.1 PLN e SPLNs.....	18
II.2.2 Revisão da escrita e revisores automáticos	21
II.2.2.1 Revisor <i>ReGra</i> : proposta e arquitetura.....	25
II.2.3 Avaliação de SPLNs e a mensuração do desempenho na revisão automática da escrita.	31
III. Lingüística de Corpus e a fonte de textos do ENEM.....	36
III.1 O contraponto teórico	36
III.2 Contexto – a era da informática.....	42
III.3 Delinamento do campo	46
III.3.1 Abordagem Baseada-em-Corpus	49
III.3.2 Abordagem Direcionada-pelo-Corpus	51
III.3.3 Trabalho com corpus: uma questão de posicionamento	54
III.4 Corpus de redações do ENEM: um corpus de estudo sobre os aprendizes da língua materna	55
IV. CORVO – desenho do corpus e modelagem de aplicativos.....	67
IV.1 CORVO – o corpus	68
IV.1.1 Compilação	69
IV.1.1.1 Seleção textual.....	69
IV.1.1.2 Transposição de registros	77
IV.1.1.2.1 Digitação vs. Transposição de registro.....	77
IV.1.1.2.2 Anotação da superfície textual	80
IV.1.1.3 Armazenamento	84
IV.1.1.3.1 Nomeação de arquivos	85
IV.1.1.3.2 Estrutura do corpus	86
IV.1.1.4 Documentação	87
IV.1.1.5 Anotação	88
IV.1.1.5.1 O <i>Identificador do Desvio</i> e a anotação automática: usuário e funcionalidades	92
IV.1.1.5.2 A detecção de desvio: formação de indicadores e levantamento de dados	104
V. Análise do Corvo: reflexões e propostas.....	107
V. 1 As intervenções do ReGra e o perfil do corpus no plano ortográfico	108
V. 2 As detecções gramaticais e a tipologia de desvios	113
V. 3 A superfície textual e o processo de produção da escrita	122
V. 4 Perfil sócio-econômico dos redatores: interpretações na contramão do ENEM	128
VI. Considerações finais e perspectivas.....	135
VII. Referências bibliográficas.....	138

Índice de Tabelas e Figuras

Índice de tabelas

Tabela III-1: Composição do banco de dados do ENEM: material fonte do corpus Corvo.....	56
Tabela III-2: Faixa de rendimento das notas dos redatores do ENEM.....	57
Tabela III-3: Quadro geral do banco de dados do ENEM.....	58
Tabela III-4: Quadro geral do Base- ENEM.....	59
Tabela III-5: Quadro de formação da Base-CORVO – recortes sob dados objetivos.....	63
Tabela III-6: Base-CORVO – distribuição entre as variáveis do perfil do redator.....	64
Tabela IV-1: Distribuição em nota COMP1 da Base-CORVO.....	72
Tabela IV-2: Distribuição em origem-região da Base-CORVO.....	72
Tabela IV-3: Distribuição em nota COMP1 da Base-CORVO de A300.....	72
Tabela IV-4: Distribuição em origem-região de A300.....	72
Tabela IV-5: Distribuição de A300 por idade.....	74
Tabela IV-6: Distribuição de A300 por origem-região.....	74
Tabela IV-7: Distribuição de A300 por renda.....	75
Tabela IV-8: Distribuição de A300 por instituição.....	76
Tabela IV-9: Ilustração da anotação da superfície textual.....	83
Tabela IV-10: Tipologia de desvio para anotação do Corvo.....	90
Tabela V-1: Comparação de desempenho do ReGra (I).....	111
Tabela V-2: Comparação de desempenho do ReGra (II).....	111
Tabela V-3: Comparação de desempenho do ReGra (III).....	111
Tabela V-4: Distribuição das intervenções do ReGra.....	112
Tabela V-5: Quadro geral das detecções dos desvios gramaticais.....	113

Tabela V-6: Quadro geral das detecções dos desvios gramaticais.....	115
Tabela V-7: Quadro geral das detecções dos desvios gramaticais.....	116
Tabela V-8: Quadro geral das detecções dos desvios gramaticais.....	117
Tabela V-9: Quadro geral das detecções dos desvios gramaticais.....	118
Tabela V-10: Quadro geral das detecções dos desvios gramaticais.....	120
Tabela V-11: Detecções dos desvios gramaticais FP (sés).....	121
Tabela V-12: Frequência dos traços da superfície textual observados no Corvo.....	123
Tabela V-13: Distribuição da presença dos traços da superfície textual no Corvo.....	123
Tabela V-14: Exemplos de ocorrências no mapeamento da superfície textual do Corvo.....	126

Índice de figuras

Figura III-1: Participação das Faixas de Rendimento em cada competência na Base-ENEM.....	60
Figura III-2: Quadro de distribuição (faixa de rendimento em COMP1) da Base-ENEM congelada.....	62
Figura IV-1: Quadro de distribuição de A300 (percentual por idade).....	74
Figura IV-2: Quadro de distribuição de idade de A300 (percentual por segmentos de A300).....	74
Figura IV-3: Quadro de distribuição de origem-região de A300 (percentual por origem-região).....	75
Figura IV-4: Quadro de distribuição de origem-região de A300 (percentual por segmentos de A300).....	75
Figura IV-5: Quadro de distribuição em renda de A300 (percentual por renda).....	75
Figura IV-6: Quadro de distribuição em renda de A300 (percentual por segmentos de A300).....	76

Figura IV-7: Quadro de distribuição em instituição de A300 (percentual por instituição).....	76
Figura IV-8: Quadro de distribuição em instituição de A300 (percentual por segmentos de A300).....	76
Figura IV-9: Distribuição das notas em cada competência nas faixas de rendimento de A300.....	77
Figura IV-10: Ilustração do Texto-Fonte.....	78
Figura IV-11: Modelo de nomeação de arquivos do Corvo.....	85
Figura IV-12: Estrutura geral do Corvo.....	86
Figura IV-13: Modelo de cabeçalho do Corvo.....	87
Figura IV-14: Ilustração de IDv – apresentação e funcionalidades (I).....	93
Figura IV-15: Ilustração de IDv – apresentação e funcionalidades (II).....	95
Figura IV-16: Ilustração de IDv – apresentação e funcionalidades (III).....	96
Figura IV-17: Ilustração de IDv – apresentação e funcionalidades (IV).....	98
Figura IV-18: Ilustração de IDv – apresentação e funcionalidades (V).....	99
Figura IV-19: Ilustração de IDv – apresentação e funcionalidades (VI).....	99
Figura IV-20: Ilustração de IDv – apresentação e funcionalidades (VII).....	100
Figura IV-21: Ilustração de IDv – apresentação e funcionalidades (VIII).....	102
Figura IV-22: Ilustração de IDv – apresentação e funcionalidades (IX).....	103
Figura IV-23: Processo de detecção dos desvios e levantamento de dados.....	106
Figura V-1: Comportamento do ReGra sobre o Corvo: quantidade de intervenções.....	108
Figura V-2: Distribuição das intervenções do ReGra.....	109
Figura V-3: Quadro de distribuição de desvios e marcas de superfície.....	127
Figura V-4: Distribuição conjunta de desvios e marcas de superfície (co-ocorrência)...	127
Figura V-5: Distribuição da presença dos traços da superfície textual no Corvo.....	131

I. INTRODUÇÃO

A opção por empreender esta pesquisa foi determinada por uma trajetória de trabalho com corpus que se iniciou juntamente com o nosso primeiro contato com o revisor gramatical automático ReGra¹ há alguns anos atrás. Do ponto de vista de uma estudante de lingüística e aspirante à professora de português, foi uma experiência que, desde o início, nos pôs a observar por um ângulo bastante particular as especificidades de um texto contendo “problemas de escrita”. Este trabalho, assim, representa a determinação que tivemos em certo momento de investigarmos cuidadosamente esse tipo de texto. Procurando aliar a experiência que desenvolvemos ao longo dos anos como professora atuante no ensino fundamental em diante com a experiência adquirida em longa temporada no NILC², o laboratório de desenvolvimento do ReGra, o que trazemos neste texto é a consolidação de algumas idéias acerca de corpus de textos não-revisados e suas implicações no desenvolvimento de um revisor como o ReGra.

No início de nossa experiência com corpus no NILC, em que compilávamos as amostras de seu primeiro “corpus não-corrigido”, digitando efetivamente os textos, o conhecimento que se tinha no Brasil sobre a importância e as particularidades de um trabalho com corpus era pequeno. Ao longo de 10 anos, porém, esse cenário foi alterado substantivamente; hoje, não é possível ignorar a presença de importantes discussões sobre corpus tanto no universo da lingüística, em que se incluem os estudos do português, quanto no universo do processamento de língua natural, de que faz parte a revisão automática.

Com efeito, isso inspirou o modo como abordamos o tema do corpus neste trabalho. Os capítulos III e IV, por exemplo, trazem um repertório que, por ter sido muitas vezes reclamado por nós, que trabalhamos efetivamente com corpus e revisão automática, merece ser apresentado com vagar, contextualizando-o, buscando revelar os interesses particulares desse campo interdisciplinar pelos textos não-revisados.

¹ *ReGra* = Revisor Gramatical Automático (do português brasileiro escrito)

² *NILC* = Núcleo Interinstitucional de Lingüística Computacional (www.nilc.icmc.usp.br)

Ao longo desta exposição, o trajeto perseguido será o de mostrar o tema do “erro” ou, como preferimos, *desvio da norma padrão* sob a perspectiva da valorização, isto é, de que a ocorrência de um desvio não revela simplesmente um “problema” da “boa escrita”, mas, antes, um indício do processo de assimilação da escrita culta pelo falante nativo. Em se tratando de análise lingüística, baseada em corpus, partimos do processo de consolidação do próprio termo “corpus” e a influência do amadurecimento da Lingüística de Corpus (LC) nesse processo, para reunir os fundamentos teóricos que sustentaram a nossa iniciativa de construção de um corpus efetivamente – o *Corvo*. Em meio a tudo isso, trazemos o revisor ReGra como um material importante na metodologia de pesquisa, empregado na expectativa de que a preparação do corpus se desenvolvesse o máximo possível pela via automática.

Podemos dizer que nos posicionamos como *projetista* de um corpus. De certo modo, buscamos construir a primeira parte deste trabalho (os capítulos teóricos) sob a influência direta do trabalho com o Corvo. Agimos como um projetista no sentido de que apresentaremos, por exemplo, o repertório da LC efetivamente requerido quando da modelagem de um corpus e não os princípios abrangentes e às vezes polêmicos que interferem na consolidação dessa grande área de conhecimento. É também como um projetista que nos dispomos a desenvolver aplicativos computacionais para servir à preparação e manipulação de um tipo especial de corpus – o corpus de aprendiz.

Nosso leitor conhecerá, desse modo, a visão teórica com a qual nos vinculamos para abordar a questão da norma padrão, dos desvios e dos textos não-revisados. Trazemos no capítulo II um repertório sobre a escrita (padrão culta) sob a perspectiva atual da variação e mudança lingüísticas. Abordamos questões ainda hoje controversas, como a definição de uma norma padrão do português, mas também circunscrevemos essa temática ao contexto do ensino e da avaliação de desempenho. A noção de desvio, neste capítulo, é desenvolvida gradualmente, à medida que alguns preconceitos lingüísticos são neutralizados e a tão verberada falência do português é desmistificada.

Por pertinente no curso da metodologia e análise do corpus, também nesse capítulo apresentamos alguns conceitos e especificidades da revisão da escrita, mais exatamente, a revisão automática. Nesse contexto, apresentamos detalhes do sistema de revisão automático aqui empregado – o revisor ReGra – de modo que, posteriormente, nosso leitor possa compreender de

que modo os resultados da nossa pesquisa podem interferir no desenvolvimento (ou na avaliação) desse sistema. O ENEM, grande referência nesse trabalho, é também desde já abordado. Mas ele é, na verdade, um assunto que permeia toda a exposição.

Está de volta, por exemplo, no capítulo III, onde enfatizamos os fatores que levam um lingüista de corpus a se dirigir à análise empírica para descrever ou explicar fenômenos da linguagem, por exemplo, os desvios da norma padrão presentes em textos do ENEM. Tem-se ainda a relevante discussão sobre importância do computador nesse campo dos estudos empíricos da atualidade. Também é nesse capítulo que trazemos algumas especificidades do *design* de corpus, tal como é definido pelos padrões internacionais, de modo que se compreenda, posteriormente, determinadas decisões tomadas na construção do corpus Corvo.

Com esse repertório técnico partimos para a exposição dos materiais e métodos empregados na pesquisa. O capítulo IV apresenta, passo a passo, o processo de construção do Corvo, em que se inclui o trabalho estatístico sobre os dados do ENEM, a utilização do revisor ReGra na chamada “anotação” do corpus, i.e., a inserção de etiquetas para indicar determinadas propriedades lingüísticas dos textos (no caso, etiquetas de desvios). Mostramos, ademais, nossa própria ferramenta de anotação – o Identificador e Editor de Desvios –, que assume, como veremos, importante papel no modo como conduzimos a investigação.

No capítulo V temos, enfim, a análise de corpus. As conclusões a que chegamos com a manipulação dos textos e o cruzamento dos dados, tanto lingüísticos como de perfil de redatores, demonstrarão fatos que caracterizam esse tipo especial de texto, de redator e de instrumental de investigação.

As perspectivas para pesquisas futuras a partir do que apresentamos aqui são apontadas na seção final. Certamente não só o ENEM, mas todos aqueles que lidam com textos não-revisados continuamente produzidos no país podem se beneficiar do nosso trabalho, mas será preciso ter clareza dos limites e dos ingredientes específicos desta pesquisa. Será preciso refletir! E se alcançarmos esse feito, teremos atingido o êxito maior da nossa iniciativa.

II. Desvios da norma padrão e revisão lingüística automática

II.1 Das noções de *norma padrão* e de *desvio*

Considerada um dos bens mais preciosos e um legado que se perpetua em todas as gerações do ser humano a escrita é a manifestação que leva o sujeito-autor a um comportamento lingüístico padrão de maior rigor. Em qualquer língua natural, produzir um texto escrito que satisfaça as expectativas do próprio redator e de seu interlocutor, ou seja, que estabeleça a comunicação, não basta. Há enquadramentos de diversos tipos, o principal deles à modalidade de língua (ou *variante*) em pauta no ato da comunicação. Segundo Rodrigues (2004):

Não há língua que seja, em toda a sua amplitude, um sistema uno, invariado, rígido. Ainda que freqüentemente se defina cada língua como *um* sistema de comunicação e os métodos de análise e descrição lingüística sejam delimitados em geral a partir do pressuposto de que se opera com uma estrutura bem determinada, sabemos que isso resulta de abstração feita conscientemente a fim de possibilitar um mais imediato domínio da estrutura lingüística por parte do investigador. Na realidade, toda língua, quer sirva a uma grande nação consideravelmente extensa e muito diferenciada cultural e socialmente, quer pertença a uma pequena comunidade isolada de apenas poucas dezenas de indivíduos, é um complexo de variedades, um conglomerado de variantes. (RODRIGUES, 2004, p. 11, grifo do autor)

Ainda segundo Rodrigues, há variações lingüísticas motivadas por diversos fatores, o que nos permite falar em variantes espaciais, variantes de classe social, variantes de grupos de idade e as de grupos de sexo, além das variantes de gerações (diacrônicas). E também tem as variantes de grau de formalismo e as variantes de modalidade, tidas como variantes de “segunda ordem”, recentemente denominadas *registros* (de língua).

Os registros se manifestam de forma oral e *escrita*, como é consenso geral. Mas, para nós, também se inclui aqui o registro eletrônico, fruto da era digital em que vivemos desde o último século, quando se desenvolveu uma modalidade de língua em que os *usuários* se manifestam em meio eletrônico, computacional, construindo as especificidades da modalidade lingüística.

A escrita é um registro de língua ancestral e que só ocorre em sociedades letradas. Por uma diversidade de fatores, na escrita há uma escala de *formalidade* que lhe é bastante característica. Podemos falar de grupos de textos menos ou mais formais, mais enquadrados em função de uma ou outra condição do discurso, uma ou outra regulação que circule pela comunidade de redatores. Um texto escrito formal invariavelmente é regulado por sua *norma lingüística* – uma espécie de “proceder” aceito entre os membros dessa comunidade. De acordo com Faraco (2004),

(...) os grupos sociais se distinguem pelas formas de língua que lhes são de uso comum. Esse uso comum caracteriza o que se chama de a *norma lingüística* de determinado grupo. (...) Como a respectiva norma é fator de identificação do grupo, podemos afirmar que o senso de pertencimento inclui o uso da forma de falar característica das práticas e expectativas lingüísticas do grupo. Nesse sentido, a norma, qualquer que seja, não pode ser compreendida apenas como um conjunto de formas lingüísticas; ela é também (e principalmente) um agregado de valores socioculturais articulados com aquelas formas. (FARACO, 2004, p. 39, grifos do autor)

Esse “senso de pertencimento” de que fala o autor é verificado no comportamento das comunidades lingüísticas em que determinadas generalizações compõem um padrão (cultural) de língua. Na *língua padrão* generaliza-se um modo de comportamento lingüístico tido como ideal, em que se observa uma tendência dos membros da comunidade pela aceitação de suas normas, embora “nem todos cumpram ou não cumpram integralmente”, como lembra Rodrigues (2004, p. 13). Para o autor, “trata-se de uma variante [de língua] tida, pelos membros de uma sociedade, como aceitável em determinadas situações, em que outras variantes podem não ter a mesma aceitação.” (RODRIGUES, 2004, p. 14).

No registro escrito padrão, ou língua escrita padrão, o que se estabelece como aceitável compõe aquilo a que se denomina *norma padrão*. Para Castilho (2004), o que determina essa aceitação é o prestígio, aspecto que também afeta um outro tipo de norma estritamente ligada à padrão, que é a *norma prescritiva*. Diz o autor:

Norma objetiva, explícita ou *padrão* real é a linguagem efetivamente praticada pela classe social de prestígio, que se podia identificar no Brasil de hoje como a chamada classe social culta, escolarizada. Trata-se de um dialeto social que em si nada tem de “melhor” em relação aos demais, decorrendo seu prestígio unicamente da importância da classe social a que corresponde. (...) *Norma prescritiva*, [por sua vez, é aquela para quem]

merecem ser ensinados os usos lingüísticos de uma classe prestigiosa considerados mais adequados a cada situação e melhor identificados como o ideal de perfeição lingüística. (CASTILHO, 2004, p. 30, grifos nossos)

Nem a norma padrão nem a norma prescritiva se confundem com a chamada *norma culta*. Na concepção desta última, tal como observa Faraco (2004), está o fato de que ela é uma regulação estabelecida estritamente pelos membros que lidam direta ou indiretamente com a “cultura” da escrita. Nas palavras do autor,

(...) [a expressão norma culta] não se confunde com norma-padrão. (...) a *norma culta* é a norma lingüística praticada, em determinadas situações (aquelas que envolvem certo grau de formalidade), por aqueles grupos sociais mais diretamente relacionados com a cultura escrita, em especial por aquela legitimada historicamente pelos grupos que controlam o poder social. (...) Seu posicionamento privilegiado na estrutura econômica e social os leva a se representar como “mais cultos” (talvez porque, historicamente, tenham se apropriado da cultura da escrita como bem exclusivo, transformando-a em efetivo instrumento de poder) e, por consequência, a considerar a sua norma lingüística – mesmo difusa em sua variabilidade de pronúncia, vocabulário e sintaxe – como a melhor em confronto com as muitas outras normas do espaço social. (FARACO, 2004, p. 40, grifos do autor)

Podemos dizer, à luz do autor, que a cultura da escrita, quando se associou ao poder social ao longo da história, desencadeou um processo unificador, visando a uma “relativa estabilização lingüística, buscando neutralizar a variação e controlar a mudança” (FARACO, 2004, p. 40). É justamente nesse ponto que norma culta e norma padrão andam juntas: a referida estabilização lingüística é realizada pela norma padrão, satisfazendo os desejos da comunidade lingüística culta. A norma culta, nesse sentido, é uma decorrência cultural da norma padrão.

A formação da norma padrão do português culto, especialmente o português escrito, foi afetada pelas inclinações político-culturais do Brasil desde a colonização. O processo que se deu a partir do século XIX desencadeou a codificação de uma norma padrão artificial, já que a elite letrada e conservadora do Brasil à época estabeleceu a literatura (Romântica) lusitana como modelo de codificação, impedindo (ou procurando neutralizar) qualquer tentativa de abasileiramento dessa norma.

Muito cedo, como se vê, a norma padrão do português se distanciou da norma culta. Efetivamente, hoje, o que se tem é uma norma culta em uso pela comunidade letrada (mas não só ela), especialmente na língua oral, e uma norma padrão em grande medida desconectada desse

uso; a norma prescritiva, se a incluirmos nessa reflexão, diz-se mais distante ainda da norma culta, já que os manuais normativos do português, ainda mais conservadores, estabilizam um uso lingüístico muitas vezes obsoleto, não praticado mesmo entre os membros tidos como mais cultos da comunidade letrada.

II.1.1 Referências da norma padrão no Brasil e a “falência” do português culto

Há um sentido, certo significado sociolingüístico da norma padrão. Ainda que artificial, como no caso do português brasileiro, em comunidades lingüísticas com vastas dimensões como a nossa é positivo o aspecto da uniformização lingüística especialmente no caso da língua escrita culta. Conforme observa Faraco (2004):

(...) a função básica de um padrão de linguagem é estimular, pelo menos na escrita, uma relativa uniformização lingüística num amplo e diversificado espaço sociocultural. Uma tal uniformização, mesmo que sempre relativa, tem indiscutível relevância em sociedades do porte como a nossa, no sentido de garantir uma base de comunicação supra-regional, transtemporal e multifuncional; ou, como diz Haugen (2001 [1966]), aquele ideal de língua caracterizado pela variação mínima na forma e máxima da função (sociocultural). (FARACO, 2004, p. 47)

O problema está em admitir que a norma padrão do português deva se manter longe do uso da norma culta e severamente resistente às questões de mudança e variação lingüísticas.

Durante muito tempo, imperou no país a idéia de que nos compêndios gramaticais, popularizados pelos gramáticos mais conservadores, a norma padrão estaria reunida. A norma prescritiva, por sua vez, baseou-se nesses mesmos compêndios, difundindo entre os *aprendizes* da norma culta um saber que cada vez mais mostrou-se destoante da realidade lingüística dos falantes do português culto para a qual estavam sendo preparados. É bastante recente, pois, o surgimento de gramáticas diferenciais como fonte de referência inovadora da norma padrão do português.

Com os estudos lingüísticos empíricos houve certa alteração nesse cenário tradicional e ortodoxo e passamos a notar a presença (e influência gradual) de gramáticas mais flexíveis, como as de Celso Cunha e Lindley Cintra, onde se documenta fenômenos e tendências da norma culta, num movimento que procura desencadear a incorporação desses fenômenos à norma/língua

padrão. Na verdade, tal como assinala Castilho (2004), o que parece ter havido é uma tal conscientização de que tanto a norma padrão quanto a norma prescritiva “não [estão] a salvo do fenômeno da variabilidade lingüística” (idem, ibidem, p. 30).

Com efeito, resistir à variação e mudança lingüísticas é favorecer um entendimento enganoso de que o português culto está falido, já que *não se escreve mais como antigamente*. Julga-se, pois, nos dias de hoje, que a norma padrão está permissiva e a norma prescritiva, impossível de ser apreendida, quiçá aplicada pelos aprendizes no uso da língua culta padrão. O que há, contudo, não é a performance fracassada do aprendiz, do redator do português culto, mas, em grande medida, uma norma padrão e norma prescritiva de difícil aceitação e convivência. Conforme lembra de modo muito pertinente Castilho, o que se processa é uma busca pela “autenticação” da norma sobre bases remotas, em franco descompasso com o uso a que correspondem:

Dado que as aspirações e atitudes mudam com o tempo, parecendo haver uma correlação mobilidade social e enfraquecimento da norma e ao contrário, estagnação social e enrijecimento da norma, é forçoso concluir que a autenticidade e a veracidade da norma radicam no contemporâneo. Com isso, não se devem buscar seus contornos em épocas passadas. (CASTILHO, 2004, p. 31)

As gramáticas tradicionais, norteados pelo ideal de norma, disseminaram entre os usuários e os aprendizes da língua culta padrão o binômio estereotipado do *certo* e *errado*. Estar em conformidade com as expectativas da escrita culta passou a significar produzir um texto correto, de onde, como é muito freqüente na atualidade, advém a percepção simplista de que *o brasileiro escreve errado*, porque não se mostra aplicado no aprendizado da norma padrão. O que se produziu no país com o tempo e com esse tipo de estereótipo foi a visão impressionística de crise permanente do português, de fracasso do ensino da gramática e de incapacidade de expressão dos indivíduos na língua culta. A questão parece ser, porém, que não se questiona a imobilidade das gramáticas tradicionais, que permanecem como fontes primárias de uma norma padrão, como vemos, cada vez menos identificada e capturada na variante de língua a que corresponde.

A dicotomia correto-incorreto, certo-errado está centrada no entendimento arbitrário de certos setores da sociedade, segundo o qual somente o conhecimento das regras normativas e a leitura dos escritores consagrados na língua culta conduzem o indivíduo ao bom desempenho da escrita. Em vista disso, surge com o tempo o movimento de denúncia da decadência do idioma,

do mau uso do vernáculo ou da degeneração da língua culta. A mídia brasileira, em especial a imprensa jornalística, passou do tom da denúncia para adotar também a postura da prescrição, elaborando manuais de redação dispostos a *sanar o mal* que assola o país na questão lingüística.

De acordo com Chociay (2004), pode-se depreender desse cenário dois discursos que nascem juntos e permanecem intactos nos dias de hoje: aquele que o autor chama de “retórica da perdição”, em que se atesta o fracasso do português culto, da norma padrão e, por tabela, do ensino e do aprendizado da língua culta, e a “retórica da salvação”, em que a proposta é a cura para esse mal empreendida pelos mais diversos setores salvacionistas que incluem, além dos manuais de redação dos grandes jornais do país, também “os serviços telefônicos e televisivos para ensinar o que é certo e errado, os serviços de plantão de língua portuguesa veiculados na internet”. Ainda segundo o autor, não bastasse a extrapolação de suas funções, esse setor da sociedade também passou a investir-se do papel de postular

(...) metodologias salvadoras, algumas delas resultantes de falsos postulados pedagógicos, outras oriundas de teorias lingüísticas mal digeridas, outras ainda nascidas de meros palpites: ensinar sem apelar a qualquer gramática; retornar aos métodos dos retóricos gregos e latinos; reformar a escola em todos os seus níveis; remunerar melhor os professores; ensinar o povo a amar o vernáculo; transformar os meios de comunicação em veículos de ensino de língua e de redação; limitar a influência das línguas estrangeiras sobre a portuguesa; exterminar as gírias e os calões, eliminar a gramática normativa; substituir a Nomenclatura Gramatical Brasileira por outra mais científica; introduzir conteúdos e práticas de Lingüística Textual, etc., etc., etc. (CHOCIAIY, 2004, p. 19/20)

Na verdade, como aponta Faraco, esses manuais de redação, que se tornam referências modernas (*da* norma padrão e *para* a norma prescritiva), “nada mais fizeram do que transcrever acriticamente o que está estipulado nos velhos compêndios da gramática” (FARACO, 2004, p. 47). E a questão retoma, assim, o ciclo vicioso.

Felizmente, conforme Mattos e Silva (2004), os estudos da variação e mudança lingüísticas conseguem a esse mesmo tempo alguma forma de penetração no debate sobre as norma padrão e culta, a construção de fontes de referência e o ensino. Principiam esvaziando a rigidez das gramáticas na medida em que se vêem introduzidos fenômenos lingüísticos reconhecidos no uso da língua culta, reformulando o posicionamento e a função que devem ter os professores na transmissão do saber sobre a norma culta e assim por diante. Teve início, enfim,

verdadeiras iniciativas de “ajustes” em favor da desmistificação junto à sociedade civil da falência do vernáculo.

Enfocando a noção do certo/errado e o modo como ela vem se transformando com o passar do tempo em virtude dessa nova influência conceitual, Mattos e Silva (2004) indica qual é o elemento inovador e a conseqüente alteração que se faz necessária no ensino (da gramática) do português culto padrão:

(...) a questão histórica no ensino da língua materna, a do certo e do errado – conseqüência do chamado “erro clássico” da tradição gramatical (Lyons, 1979), arejada pela aceitação de postulados da sociolinguística e das teorias sobre práticas do discurso, que se refletem na substituição dos qualificadores tradicionais por outros menos coersitivos, o adequado e o inadequado –, implica um ensino de “gramática” que terá por objetivo ajustar às normas de uso idealizado como “corretas/adequadas” os usos diversificados do falante nativo antes livre da instituição normativizadora. (MATTOS E SILVA, 2004, p. 81)

Ora, por essa perspectiva, falamos não mais de certo e errado, mas de *adequado* e *inadequado*. Como a própria autora se expressa em diversos momentos, trata-se de se referir ao inadequado não mais como um “erro”, antes como um *desvio*, especialmente no que diz respeito à inadequação do indivíduo ao registro escrito formal em sua variante culta padrão.

Um expressivo número de lingüistas contemporâneos concorda que essa mudança de nomenclaturas não é mero ajuste técnico. Ao contrário, opera em conjunção com uma tendência (inclusive social) de se interpretar o ato da escrita como uma espécie de *posicionamento consciente* do usuário da língua frente a situações discursivas específicas. Por trás da noção de desvio, por oportuno, está a compreensão de inadequação a uma variante de língua que é uma dentre tantas, que é peculiar e regulada, e cujo sucesso depende de aprendizado formal (da escola) absolutamente eficaz.

II.1.2 Desvios da norma padrão e a mensuração do desempenho do redator

Sem, portanto, se definir de modo consensual qual é a norma padrão vigente, os redatores usuários da língua culta bem como os aprendizes dessa variante no país têm sido submetidos, há décadas, a avaliações de seu desempenho na escrita formal. Fortemente notadas no âmbito

escolar, essas avaliações assumem de forma retórica que os estudantes são satisfatoriamente expostos ao conteúdo e ao treinamento da norma/língua culta, desenvolvendo competência para o uso da norma padrão durante a formação escolar (de aproximadamente 15 anos após o período de letramento).

A produção escrita treinada em ambiente escolar está concentrada na elaboração de um texto conhecido como *redação*. De acordo com Lara (1994):

[foi] na segunda metade da década de 70 [que se deu] o início do processo de redemocratização e a conseqüente crítica da ideologia subjacente à utilização da tecnologia instrucional [que] restituíram a palavra ao estudante. Nessa mesma época, o decreto de no. 79.298, de 24.02.77 (...) determinou a volta da redação à escola através da “inclusão obrigatória da prova ou questão de redação em língua portuguesa” nos concursos vestibulares (Art. 1º, alínea d). Assim, para fazer face à nova exigência, a escola passou a enfatizar a produção de textos nas aulas de língua portuguesa, chegando até mesmo a criar disciplinas com o objetivo específico de ensinar a redigir. (LARA, 1994, p. 66)

Tal como assinalamos atrás, ao lado desse movimento oficial, houve uma mobilização da imprensa que, a partir do mesmo processo de redemocratização citado por Lara acima, insistiu na denúncia de uma suposta “incapacidade de expressão escrita dos estudantes (principalmente os vestibulandos)” (LARA, 1994, p. 67). Ainda segundo a autora, foi a partir desse conjunto de fatores que não só a produção escrita dos aprendizes passou a ser de interesse dos pesquisadores no país, como também passou a pertencer aos fatos observados na formação escolar desse indivíduo, incidindo sobre a redação um sistema de mensuração do desempenho do redator estritamente relacionado ao uso da norma padrão.

Ao longo do tempo, e à guisa dos inúmeros descontentamentos e confrontos por eles motivados numa parcela da sociedade, as redações transformaram-se em itens essenciais dos concursos vestibulares, de processos de seleção para ingresso no mercado de trabalho e, mais recentemente, de alguns mecanismos do sistema de avaliação em larga escala do país (p.ex.: SAEB e ENEM)³. Em vez de empenhados em desatar os nós da norma padrão, no que diz respeito à avaliação da produção escrita dos aprendizes, o que se observou com esses mecanismos, porém, foi o reforço dos estereótipos de que já falamos aqui, assumindo de modo muito vago que existe um *domínio da norma culta* (na verdade, norma padrão) no qual o aprendiz

³ SAEB = Sistema Nacional de Avaliação da Educação Básica; ENEM = Exame Nacional do Ensino Médio.

egresso do ensino médio, no caso do ENEM, tem seu texto escrito pontuado quanto ao grau e a eficiência do uso das regras prescritivas, supondo que o ensino de português, dedicado a esse tema em quase todo o período de escolarização, lhe tenha fornecido as bases para o sucesso na avaliação.

Queremos encerrar esta exposição teórica em torno da norma padrão, trazendo um posicionamento crítico que surge a este tempo no país, através dos quais se questionam, justamente, os critérios e a validade das medidas de avaliação nesse quesito da norma culta em redações produzidas, especialmente, pelos candidatos do ENEM.

Na contramão do entendimento de que o aprendizado eficiente da norma padrão se reflete no uso efetivo da língua culta, o ENEM toma em separado o desempenho do redator no quesito norma culta, transformando-o em algo passível de mensuração. Há, por paradoxal, um discurso e uma prática divergentes nas diretrizes do ENEM, para as quais, em pressupostos técnicos, pode-se medir a capacidade dos redatores de operar com o padrão, ao mesmo tempo em que, na prática da avaliação, o que se conta é a quantidade de conhecimento enciclopédico da norma padrão que esse redator apresenta em sua redação.

Segundo Faraco (2004), para quem o ENEM se configura mais um mecanismo de reforço dos estereótipos da língua culta sem vantagem alguma para a inovação tão urgente da norma padrão,

(...) aquelas diretrizes [do ENEM], ao inverterem o discurso pedagógico e ao isolarem a norma-padrão, apenas reiteram, sob um manto de aparente inovação, a clássica reificação e fetichização escolar da norma-padrão. O produto maior da escolaridade média na área da língua não pode ser o domínio de um objeto recortado no abstrato (como tradicionalmente se faz na escola brasileira), mas – como temos intensa e extensamente debatido há décadas – o domínio de práticas sociais próprias da cultura escrita, no interior das quais (e só aí) faz sentido falar de norma-padrão. (FARACO, 2004, 57)

Não vem ao caso, embora seja importante, revelar e se aprofundar nas tais diretrizes do Exame, mas relacionar essa iniciativa de avaliação ao impacto que tem sobre o uso da norma padrão, tal como vimos abordando até aqui. Pensemos, afinados com Scherre (2004), que mais uma vez a língua culta que se produz no Brasil sofre com o flagelo em que se encontram as instituições de ensino, principalmente públicas, que deveriam tomar as rédeas da situação e

transformar o ensino da gramática (ou da norma padrão) em algo que extrapolasse o saber enciclopédico abstrato, porque desconectado das práticas sociais pertinentes. Em outras palavras, seria mais proveitoso ensinar a diversidade típica da língua materna, procurando tornar os aprendizes sujeitos capazes de transitar pelas variantes de língua, manipulando os recursos lingüísticos normatizados em favor de sua maior aceitação social no universo letrado, culto. Tal como enfatiza Scherre,

Podemos atualizar e ensinar gramática normativa sem menosprezar a língua materna, nas suas diversas manifestações dialetais. Repito: o domínio da língua materna – entendida como primeira língua – é natural. Não requer ensino. O domínio de gramáticas normativas, de segundas línguas, de línguas estrangeiras, de processos de leitura e de processos de escrita, este, sim, é adquirido. E, para ser adequadamente adquirido, requer ensino eficiente e exercício exaustivo. (SCHERRE, 2004, p. 245)

O que tem feito o ENEM, ainda que não seja uma propriedade dele exclusiva, ao conferir destaque à medida do desempenho dos aprendizes no quesito norma padrão e extrair dali indicadores classificatórios dos redatores-candidatos, é intensificar a já desoladora situação do português culto. Conforme lembra Scherre,

A exclusão pela linguagem é certamente um dos maiores fatores de exclusão social. Por ser a língua um poderoso mecanismo de identidade, mas um perverso instrumento de poder, a sociedade transita pela fronteira entre a identidade e o poder às vezes sem perceber e, assim, corrobora, inexoravelmente através dos tempos, todas as abomináveis práticas de preconceito lingüístico. (SCHERRE, 2004, p. 247)

No nosso trabalho, essa postura de desmistificação do ENEM passará da simples crítica à tentativa de mostrar que algo mais proveitoso com as redações é possível. Veremos, no decorrer da exposição, de que modo os pressupostos teóricos que trouxemos até aqui se mostrarão na opção pelos métodos e materiais de investigação.

II.2 Das noções de *desempenho* e de *revisão automática*

De todo modo, se a idéia de mensuração do desempenho não seduz ou pouco crédito tem na relação norma padrão-ensino, quando o registro escrito se realiza em ambiente digital, mensurar o desempenho de um redator torna-se uma prática pertinente e necessária. Para quem

busca entender e processar uma língua por meio do computador, dos sistemas computacionais ligados à escrita, a norma padrão é levada em conta especialmente porque, nesse ambiente, o sujeito revisor, aquele que julga o grau de adequação lingüística de um texto, não é um sujeito humano, mas artificial. O quanto esse sistema de processamento lingüístico sabe sobre a norma padrão do português? Por que a norma padrão aqui é pertinente? Em quais sistemas elas estão presentes?

II.2.1 PLN e SPLNs

Indagações como essas são colocadas nas pesquisas em *PLN* – Processamento de Língua Natural. Segundo a definição genérica de Joshi (1991),

O processamento de língua natural (PLN) é o estudo dos modelos matemáticos e computacionais de diversos aspectos da língua bem como o desenvolvimento de um amplo conjunto de sistemas. (...) O PLN tem um papel bastante especial, porque muitos campos da ciência computacional lidam com aspectos da computação lingüística, e o PLN objetiva modelar a língua computacionalmente. (JOSHI, 1991, p. 1242)⁴

Essa definição contempla dois aspectos significativos do PLN: o de que estamos na esfera da “modelagem” computacional utilizando-se a língua natural como o elemento de comunicação, e o de que, para se obter algum sucesso, é necessária compreensão ampla e integrada dos diversos fenômenos lingüísticos por parte dos “desenvolvedores de sistemas”.

Ao sublinharmos esses dois aspectos, queremos destacar o fato de que, em PLN, uma certa descrição lingüística transforma-se em “modelo” de língua para a máquina operar satisfatoriamente, subsidiando a construção de “aplicativos”, cuja eficácia pode ser medida pela cautela e ousadia dos projetistas ao tratar a língua natural. Mas o amadurecimento desse tipo de intercâmbio entre lingüistas e informatas (os desenvolvedores ou projetistas de sistemas), tão importante, deu-se com alguns embates.

Historicamente, o PLN acumulou fortes polêmicas com a lingüística teórica, quando a idéia de desenvolver programas computacionais inteligentes revelou a importância do estreito

⁴ “Natural language processing (NLP) is the study of mathematical and computational modeling of various aspects of language and the development of a wide range of systems. (...) NLP has a very special role because many branches of computer science deal with linguistic aspects of computation and NLP aims to model language computationally.”

compromisso com teorias e descrições lingüísticas de alto nível. Essas, porém, quando existiam, ou não eram consensuais ou se apresentavam em metalinguagem inacessível para os informatas.

Conforme Dias-da-Silva (1996), como efeito colateral indesejável, constatou-se certa lentidão de iniciativas de pesquisa e desenvolvimento; em contrapartida, tal embate mostrou-se salutar não só para a lingüística, com maior abertura para as estratégias de investigação do PLN, como também para o PLN, conseguindo escapar de modelos imediatistas, sem fundamentação lingüística satisfatória.

O entendimento que vigora nos tempos atuais é o de que os modelos lingüísticos de base num programa em PLN, consideradas as limitações computacionais gerais, devem responder satisfatoriamente sobre a competência e o desempenho de um “usuário de língua” nativo, pois somente assim desencadearão estratégias de processamento mais realistas. Tal entendimento é verificado nos propósitos de um “programa” em PLN.

Segundo Dias-da-Silva (1996), os grandes propósitos do PLN estão contidos em certos programas computacionais, ou mais especificamente nos Sistemas de PLN – os *SPLNs*. Os *SPLNs* são programas que, em última instância, procuram aproximar a máquina do universo lingüístico humano, na medida em que abandonam as instruções e comandos codificados em linguagem artificial e adotam a língua natural em seu lugar. Do ponto de vista do desenvolvedor:

Projetar SPLN significa, antes de tudo, tornar os computadores máquinas mais acessíveis, principalmente ao usuário comum (...) Criar programas que facilitem a comunicação entre o computador e o usuário, já iniciado no universo da informática, ou não, significa, portanto, desenvolver sistemas computacionais que incorporem algum tipo de SPLN, isto é, um conjunto de programas específicos, integrado ao sistema e projetado para executar a complexa tarefa de interpretar e gerar informações veiculadas por mensagens lingüisticamente construídas. (DIAS-DA-SILVA, 1996, p. 66)

Por tal direcionamento, o projetista de PLN tem diante de si um grande desafio. Primeiro, em função do conhecimento interdisciplinar necessário à construção de sistemas; um conhecimento que deverá absorver e tornar operacionalizável. Segundo, porque, via de regra, um *SPLN* não é criado para agir só; em vez disso, projeta-se algo preparado para ser articulado a outros sistemas e ser incluído num ambiente maior de trabalho. Em síntese, o projetista logo percebe que participa de algo maior, que é a aplicação.

Tratando do tema do léxico e sua função de prover determinados SPLNs, Santos (2000) demonstra que, até recentemente, os desenvolvedores de SPLNs careciam de projeção quanto ao seu pertencimento nos debates e estudos lingüísticos. Para reverter isso, foi preciso chamar a atenção para os procedimentos de uma certa metodologia de investigação, que dificultava a ascensão do projetista. Tornou-se primordial, então, reconhecer no projetista um pesquisador que, pela via do trabalho aplicado, tratava de maneira efetiva questões importantes do universo lingüístico. Diz a autora:

De facto, a estruturação de vários livros de introdução ao PLN parece perpetuar a idéia de que primeiro se deve estudar o léxico, os formalismos, e as teorias, e só depois pensar em aplicações. Embora haja uma certa dose de bom senso nessa forma de apresentação, gostava de poder contrariar um pouco essa tendência, chamando a atenção para a perspectiva contrária: é ao tentar resolver um dado problema (isto é, ao tentar construir um dado programa que manipula a língua) que surge o momento de nos debruçarmos sobre (algumas características) do léxico ou das teorias que o tentam resolver. (SANTOS, 2000, p. 2)

Santos afirma textualmente adiante: “estamos ‘condenados’ a aprender muito sobre a língua quando desenvolvemos qualquer aplicação.” (SANTOS, 2000, p. 15). E desse modo podemos concluir que o desenvolvimento de SPLNs contribui significativamente para a abordagem do PLN sobre as línguas naturais e não pode ser relegado ao segundo plano. O papel do projetista, a partir desse enfoque, é mais do que o de adaptar teorias; é o de se posicionar frente a elas.

Na literatura especializada, observa-se que algumas propostas de classificação das aplicações (conseqüentemente, dos SPLNs) tornam-se mais específicas à medida que o foco da discussão é a ferramenta. Vieira e Lima (2001) referem-se às ferramentas como “subsistemas”, aludindo ao fato de que o grau de envolvimento com teorias lingüísticas, formalismos, enfim a tecnologia adotada é tão elevado que não é suficiente chamá-los de ferramentas. Para as autoras, um corretor gramatical automático, que já é um sistema, possui um analisador sintático que, por sua vez, é outro sistema de processamento. Seguindo a sugestão das autoras, pois, havemos de falar em sistemas, subsistemas e também de ferramentas, sendo coisas distintas entre si, quando nos dispusermos a tratar a língua nesse âmbito do PLN. O que se depreende do debate especializado é que parece ser bastante diferente o tipo e o grau de preocupação de um projetista envolvido na construção de uma ferramenta ou, por exemplo, na construção de um subsistema.

No primeiro caso, para nos fixarmos num ponto de interesse, há uma modelagem pré-definida na qual se aposta; no segundo, as apostas podem ser de toda sorte; não há nada pré-definido no que se refere à modelagem lingüística,

Os SPLNs, em geral, são construídos para realizarem tarefas específicas e de ordem prática, o que lhes impõe a contrapartida da eficiência e precisão. Para dar conta delas, a grande maioria dos SPLNs são específicos de língua e dependentes de tarefa. Santos (2000), comentando sobre a especificidade do processamento automático do português, ressalta justamente esse atributo dos SPLNs:

Até que ponto é que há um processamento de linguagem natural, ou um processamento do português, ou do inglês? Na minha opinião, pode partilhar-se tecnologias, mas não se deve adaptar sem sentido crítico aplicações pensadas (ou implementadas) para o inglês. Não só porque cada língua tem problemas específicos (áreas que causam dificuldades diferentes a um processamento automático), mas porque soluções apropriadas a uma língua podem não ajudar muito no processamento de outra. (SANTOS, 2000, p. 4)

De fato: falamos em sistemas cujo modelo lingüístico é, por vezes, tão diferente entre si (porque, específicos de língua), que o que os aproximam é unicamente a tarefa para a qual foram projetados (ex. um corretor ortográfico). Por sua vez, encontramos sistemas cujo modelo lingüístico é tão genérico, que pode servir para desempenhar diferentes tarefas (ex. tokenizar e contar frequências).

De todo modo, não restou dúvidas, com o tempo, de que o desenvolvimento de SPLNs é uma maneira de se estudar uma língua natural e que, apesar de certa eficácia comprovada, os SPLNs não são genéricos, mas específicos ou *dependentes de língua e/ou de tarefa*.

II.2.2 Revisão da escrita e revisores automáticos

Sistemas de apoio à escrita são SPLNs com o propósito de auxiliar o usuário na redação de um texto em língua natural num ambiente de edição eletrônica. De modo geral, esses sistemas são específicos de língua (p.ex., do português brasileiro) e têm como objetivo oferecer ao usuário um recurso de *revisão da escrita*. Com este, espera-se que, de modo rápido e por meio de metalinguagem acessível, o texto do usuário torne-se adequado aos padrões da variante culta.

O *usuário-alvo* é um dos primeiros aspectos de importância para um revisor automático. Esses usuários são visados pela indústria da informática, que os transforma em “clientes”. Os sistemas de maior evidência, quando se trata de revisão automática, têm usuários-clientes, e o perfil do sistema é, em grande medida, determinado por eles. Há certo consenso, porém, de que os revisores automáticos são planejados para atender a usuários nativos da língua processada pelo sistema⁵.

Na verdade, os revisores são sistemas complexos e, mesmo nos dias de hoje, incompletos. Dentre suas limitações, está a incapacidade de operar plenamente em qualquer *domínio*. Diz-se, então, que os sistemas trabalham em domínio “fechado” ou “aberto”, numa oposição em que se considera aberto, por exemplo, o sistema cujo usuário seja o mais genérico. Consideremos alguns exemplos. A partir da década de 1990, os usuários dos editores Word, em diferentes línguas, foram alvo de sistemas de revisão automática, tanto ortográfica como gramatical. Temos, nesse caso, um usuário de domínio aberto, pois não é possível prever suas escolhas lingüístico-textuais, a finalidade comunicativa do texto sob revisão, e assim por diante. Outro exemplo é o usuário da internet (*webmail*, por exemplo), cujo perfil não é possível estabelecer previamente e, portanto, também se caracteriza como um usuário de domínio aberto. Como se vê, o que prevalece nesses sistemas de domínio aberto é o fato de que eles requerem determinado grau de sofisticação que talvez não se aplicasse a sistemas de domínio fechado.

Fala-se, contudo, em usuários para os quais um revisor automático deve garantir sucesso e esses são, objetivamente, aqueles que desejam se comunicar na *modalidade culta* da língua escrita. Algo que poderia soar como obviedade, mas que expressa um princípio norteador desses programas: a modalidade culta, baseada na norma padrão, se configurará como aquela para a qual esse sistema deve ser preparado para agir.

A *intervenção* do revisor no texto do usuário tem como meta resolver as inadequações aos padrões apresentadas num texto escrito, não importando se o usuário domina a modalidade culta ou se o seu texto não tem caráter formal o bastante para exigir rigor e enquadramento lingüístico. Assim, do escopo desses sistemas excluem-se tão somente os usuários que manipulam a língua

⁵ Diante de seus objetivos e desse perfil de usuário, os revisores automáticos são sistemas desenvolvidos sem o compromisso com o ensino de língua, muito embora iniciativas relacionadas ao ensino possam desenvolver-se independentemente. Knutsson et. al (2003), por exemplo, apresentam uma proposta de uso de revisor automático no ensino de língua estrangeira.

com propósito artístico, já que, para esses, a subversão aos padrões lingüísticos pode ser uma intenção.

Os sistemas de revisão automática incluem detecção e correção de *erros* e são de dois tipos: revisor ortográfico automático (*spell checker*) e revisor gramatical automático (*grammar checker*). As primeiras aplicações de auxílio à escrita fixaram-se no desenvolvimento de hifenizadores e revisores ortográficos. Hoje, os hifenizadores são aplicativos complementares embutidos nos revisores ortográficos e, cada vez mais, os próprios revisores ortográficos têm sido acoplados a sistemas maiores de revisão, que incluem a revisão gramatical.

Em termos gerais, a diferença entre o revisor ortográfico e o gramatical reside no escopo de observação do texto pelo sistema. Conforme Arppe (2000):

(...) a diferença entre um revisor gramatical e um revisor ortográfico é que, enquanto o revisor ortográfico está limitado à verificação da corretude de uma cadeia simples de caracteres entre espaços em branco, um revisor é capaz de levar em conta longas seqüências de tais cadeias, geralmente sentenças e parágrafos. Por isso, uma cadeia pode ser aceitável para um revisor ortográfico, mas identificado como errada para um revisor gramatical. (ARPPE, 2000, p. 15)⁶

Em ambos os tipos de sistemas esperam-se duas ações eficientes: a detecção do erro (subsistema “de diagnóstico”), que deve localizar o erro e apresentá-lo de forma amigável ao usuário, e a correção do erro, que acarreta a alteração propriamente dita. Nos sistemas ortográficos, existe ainda o subsistema de “críticas”, oferecendo ao usuário alternativas de alteração.

Os *revisores gramaticais automáticos* são sistemas que têm evoluído consideravelmente na última década, quer em função de novas abordagens para o tratamento da língua natural, quer em virtude da sofisticação dos seus subsistemas e recursos de apoio. Um sistema de revisão gramatical pode ser planejado para analisar a formação sintática e/ou morfológica das

⁶ “(...) the difference between a grammar checker and spell checker is that whereas a spell checker is limited to verifying the correctness of a single string of characters between two while-space characters, a grammar checker is able to take into account longer sequences of such strings, typically sentences or paragraphs. Thus, a string may be accepted by a spell checker but identified as erroneous in its context by a grammar checker.”

construções do usuário. Nesse caso, dizemos que ele contém o subsistema da análise sintática (*parser*)⁷ ou da análise morfológica.

Os *parsers* são subsistemas requisitados em outros sistemas, inclusive de outras aplicações. No sistema de revisão gramatical, ele é de fundamental importância, pelo apoio que oferece às estratégias de localização e checagem normativa do erro. Tais estratégias são essenciais para o tratamento refinado (portanto, mais eficiente) de um texto, sobretudo aqueles fora dos padrões da escrita culta.

São importantes, também, os analisadores lexicais (ou morfológicos), especialmente para a detecção e tratamento da má-formação de palavras.

Os sistemas de revisão gramatical contam, ainda, com um conjunto de *recursos de apoio*; dentre esses os mais comuns são: o conjunto de regras gramaticais de referência, o corpus e o léxico. No que diz respeito às regras, essas são, em geral, transpostas das gramáticas e de disposições oficiais para cada idioma.

Já o léxico é imprescindível num sistema desse tipo. Afinal, o apoio do léxico torna possível o maior entrosamento entre ferramenta e usuário, já que há compartilhamento lingüístico básico. Além disso, as informações adicionais contidas num léxico (p.ex., a(s) classe(s) gramatical(is) de um lexema) favorecem o refinamento do *parser* e, conseqüentemente, do sistema como um todo. Naturalmente, falamos em léxicos extensos, dinâmicos, ou mesmo de uma “base lexical”, mais complexa⁸.

Os primeiros revisores gramaticais automáticos foram desenvolvidos para o inglês. Para essa língua, desde então, a evolução dos sistemas tem sido constante e, de modo geral, constituem produtos reconhecidos e até mesmo populares, como é o caso dos revisores comercializados pela Microsoft Corporation.

Para o português, desde o início de 1990 encontramos sistemas comerciais e acadêmicos para as variantes européia e brasileira. No caso do português europeu, destaca-se o sistema *Flip* (FLIP, 2005) o qual, em 2005, encontrava-se na 5ª versão. Nessa, além das funcionalidades de

⁷ Neste trabalho, *parser* é tomado como sinônimo de analisador sintático e *parsing*, de análise sintática.

⁸ Uma base lexical geralmente contém informações não privilegiadas num léxico comum, tais como informações semânticas e/ou pragmáticas. Além disso, uma base comporta-se como um subsistema e tem mecanismos que inter-relacionam os seus componentes.

revisão ortográfica e gramatical, também se disponibilizam a revisão estilística e aplicativos associados (hifenizador, conjugador verbal, tradutor de palavras, dicionários temáticos). Uma plataforma de edição foi também desenvolvida – o FliPEd – e nela o usuário tem acesso a todos os aplicativos para auxílio à escrita produzidos pelo grupo.

Já para o português brasileiro, a mais recente iniciativa é o sistema *CoGrOO* (KINOSHITA et al., 2005). A proposta é de cunho acadêmico, envolvendo a disponibilização de um sistema de revisão gramatical automática para usuários do OpenOffice, que é um software livre (i.e., tem código aberto) e executável em diferentes plataformas, inclusive as mais comuns, como o Windows e o Linux. O CoGrOO é específico para a revisão gramatical e, por isso, aposta em novas tecnologias (como a abordagem híbrida – estatística e baseada em regras no módulo de análise sentencial) para dar conta de fenômenos complexos da modalidade escrita, como a concordância, por exemplo. Há também grande investimento da equipe no aspecto da compatibilização do sistema às plataformas do OpenOffice. Contudo, o sistema encontra-se em estado de protótipo até o momento, revelando-se ainda precário quanto à abrangência de recursos importantes, como o léxico e o corpus.

É, sem dúvidas, o revisor *ReGra* que, hoje, se apresenta como o sistema de maior projeção quando se trata do português do Brasil. Trata-se de um sistema consolidado, não só pela sua popularização, visto que está disponível no editor Word da MicrosoftTM, mas também pelo conjunto de publicações que acompanha o revisor desde a sua criação (em 1993), tornando públicas as estratégias de desenvolvimento, a metodologia adotada, bem como as limitações e os desafios desse empreendimento.

II.2.2.1 Revisor *ReGra*: proposta e arquitetura

O nascimento do ReGra aconteceu pela parceria entre a Itautec-Philco e o NILC (Núcleo Interinstitucional de Lingüística Computacional), em 1993. A empresa, na época, colocava no mercado seu editor de textos – Redator – e desejava incrementá-lo com aplicativos de auxílio à escrita. O revisor ortográfico já vinha sendo desenvolvido, mas a empresa queria a revisão gramatical, até para fazer frente à concorrência (especialmente com a Microsoft), que já apresentava esse diferencial aos usuários do editor (Word). Assim, o NILC desenvolveu um

sistema que, desde o início, delineou-se como um produto comercial, estabelecendo fortes parâmetros para as suas estratégias de construção e abordagens metodológicas.

A questão do público-alvo para o revisor foi um desses aspectos de conformidade do projeto à aplicação. No início, o usuário do ReGra foi definido como sendo secretárias, em função do perfil corporativo dos clientes da Itaotec-Philco. Então, o primeiro passo dos projetistas foi construir um corpus específico, majoritariamente composto por textos autênticos desse público.

O levantamento dos *desvios* mais freqüentes nesse corpus estabeleceu o conjunto de “problemas” com o qual o ReGra deveria lidar com prioridade. Foram eles:

✗ortografia	✗pontuação	✗influência da oralidade
✗acentuação	✗crase	✗uso inadequado de palavras

A ortografia e a acentuação geralmente são tratadas pelo revisor ortográfico e, portanto, no caso do ReGra, foram os demais tópicos que constituíram o foco inicial do desenvolvimento do revisor.

Como “pacote” de revisão da escrita, a primeira versão do ReGra e do revisor ortográfico foram comercializados no varejo a partir de 1997. Nesse mesmo ano, firma-se o acordo com a Microsoft Corporation, que passa a oferecer aos usuários do programa OfficeTM a revisão gramatical do ReGra. Por esse acordo, o ReGra substituiu o revisor do Office para o português, até então suportado por um sistema desenvolvido em Portugal, mas que não tinha abrangência, sobretudo lexical, para tratar a variante brasileira. A versão do ReGra incorporada no Office incluiu, além da revisão gramatical, também um *thesaurus* e foi lançada oficialmente no pacote do Office-2000. Por sua vez, a Itaotec-Philco prosseguiu no lançamento de novas versões do ReGra, acrescentando outros recursos de auxílio à escrita ao pacote comercial (consulta à gramática da língua portuguesa, conjugador verbal, consulta a dicionários do português ou bilíngües, etc.). Esse pacote é denominado “Redator Plus”.

A introdução da Microsoft no projeto estabeleceu novos parâmetros para o desenvolvimento do ReGra. O perfil de usuário, por exemplo, foi alterado substancialmente, deixando de ser o profissional corporativo para ser o “usuário da língua portuguesa de nível médio, com escolaridade de segundo grau completo” (BONFANTE e NUNES, 2000). A experiência com o ambiente Office, porém, demonstrou que esse público-alvo do revisor seria

mais abrangente e, assim, redefiniu-se o perfil de seu usuário, desta vez nativos da língua portuguesa do Brasil, com conhecimento mediano da estrutura dessa língua bem como da metalinguagem empregada na gramática normativa de referência. Em síntese: o ReGra, que desde o início já se voltava a um usuário aberto, ampliou ainda mais esse perfil com o novo ambiente de atuação.

O pressuposto de que esse revisor deve estar preparado para processar qualquer tipo de texto produzido por nativo do português brasileiro (i.e., processar textos de contextos não específicos) sobrelevou uma série de desafios. Dentre eles, o fato de que, efetivamente, o texto de *input* para o sistema é um texto não-revisado, que contém inadequações normativas de toda sorte que são, ao mesmo tempo, o objeto de análise do revisor e a expectativa de desempenho eficiente do usuário, i.e., o usuário espera que o revisor intervenha a seu favor. Soma-se a isso o fato de que é desejável um sistema que responda em tempo real e, preferencialmente, acerte muito mais do que erre a fim de demonstrar-se útil (e até imprescindível) ao usuário. Para se compreender essas e outras decisões de projeto e especificidades do processamento, é mister considerar a “arquitetura” do sistema.

Dá-se o nome de “arquitetura” à organização funcional de um sistema. No caso do ReGra, sua arquitetura compõe-se de três módulos funcionais e as relações entre eles: o módulo mecânico, o módulo estatístico e o módulo lingüístico.

O sistema inicia com a delimitação dos parágrafos do texto de entrada e, em seguida, com o reconhecimento e classificação dos *tokens* por meio de um analisador lexical, acionado antes de qualquer um dos módulos (veja mais adiante).

O *módulo mecânico* trata, de modo geral, dos problemas gerados por falhas de digitação (p.ex.: uso não balanceado de símbolos delimitadores, tais como aspas ou parênteses, distanciamento de caracteres, palavras ou símbolos repetidos, etc.). Os problemas que concernem o módulo mecânico, enfim, são “facilmente identificáveis e normalmente não são detectados por um corretor ortográfico” (PARDO et. al, 2000).

O *módulo estatístico* coleta medidas estatísticas do texto: número de parágrafos, de sentenças e de palavras; tamanho médio das palavras (em função das letras e sílabas), das sentenças e dos parágrafos. A partir desses dados, calcula-se o chamado “índice de legibilidade”, ou seja, uma indicação sobre o “grau de dificuldade de leitura do texto por leitores com um certo

grau de escolaridade. Dessa forma, o índice pode fornecer subsídios adicionais a uma análise da adequação do texto ao seu público alvo, e sua aplicação pode ajudar a identificar se o texto está escrito em um padrão acima do que seria adequado, dificultando, ou mesmo impedindo, a comunicação” (NUNES et. al, 1996). De todo modo, as medidas estatísticas coletadas de um texto servem para caracterizar a chamada *superfície textual* e podem ser significativas num determinado ponto da revisão automática.

Antes de partirmos para o módulo lingüístico, convém mencionar o analisador lexical. Trata-se de uma etapa fundamental do sistema, em que os *tokens* são validados em função de um dicionário⁹ especialmente desenvolvido para o revisor. Por esse processo, palavras não-existentes no dicionário são apontadas e, aquelas existentes, são recuperadas pelo sistema já com os atributos morfossintáticos.

O conjunto desses dados é utilizado durante o processo de revisão, por exemplo, na etapa de desambiguação, em que determinado *token*, morfológicamente polissêmico, é submetido à análise. O dicionário, nesses casos, oferece as categorias morfológicas ligadas ao *token* em questão, e uma indicação sobre a prioridade de uma categoria sobre a outra nos casos de polissemia.

Outra contribuição significativa desse componente lexical é a validação da boa-formação de palavras do português. O dicionário contém informações sobre prefixos e sufixos comumente associados aos *tokens*, o que é resgatado pelo sistema numa situação de verificação da formação de palavras. Enfim, esse componente, enquanto analisador, cumpre papel substancial em diferentes situações da revisão.

No *módulo lingüístico* concentram-se as etapas mais delicadas e decisivas do revisor. Segundo Nunes et al. (1996), “o módulo lingüístico é, obviamente, o módulo central do revisor” (NUNES et. al, 1996, p. 72), dado que, por exemplo, é esse componente que contempla as tarefas de análise sintática (*parsing*) e a de detecção, correção e aconselhamento gramatical (*checking*).

⁹ Empregamos o termo “dicionário” de modo simplificado neste contexto. Na verdade, e sobretudo atualmente, esse componente lexical, no NILC, apresenta-se como uma “base lexical”, contendo mais que tokens e seus atributos morfossintáticos; tem-se, p.ex., informações semântico-pragmáticas e ontológicas de lexemas, além de mecanismos fortes de associação de informações e eleição de dados. Ao nos referirmos ao dicionário, queremos especificar o componente do ReGra, que não possui a abrangência da base lexical desenvolvida no NILC.

O texto de *input* do módulo lingüístico é o resultante dos processamentos anteriores; ou seja, o texto que esse subsistema recebe está supostamente livre de erros mecânicos e desvios ortográficos, por exemplo, e está parcialmente marcado com as informações lexicais geradas na análise lexical.

O texto é, então, segmentado em unidades menores (parágrafos), em função da pontuação e marcadores textuais presentes (conjunções, p.ex.). Em seguida, tem início, propriamente, a análise sintática e a aplicação de regras de detecção e correção, numa abordagem considerada “híbrida”, devido à utilização de estratégias distintas no processamento.

Uma primeira estratégia de detecção e correção de erros gramaticais – chamada *error-driven* – atua de modo local, é sensível a contextos específicos e, por isso, é considerada uma abordagem “determinística”. Trata-se da aplicação de um conjunto de regras “pontuais”, preparadas a partir de um corpus pertinente, em que determinados “erros” são detectados como um padrão e não oferecem dificuldades para o reconhecimento no texto de *input*. O uso incorreto do verbo haver (uso do “há” em vez do “a” – preposição) é um exemplo da revisão baseada na estratégia *error-driven*. Na correção dessa inadequação gramatical, a abordagem determinística funciona com resultados bastante positivos em termos de processamento, tais como a rapidez, a especificidade e o pouco uso de memória. De acordo com Pardo et al. (2000), “a opção pelo determinismo durante a análise se deveu à necessidade de eficiência do sistema, que deve fornecer respostas em tempo real, enquanto o usuário está digitando seu texto” (PARDO et al., 2000, p. 5).

Contudo, desde as primeiras versões do ReGra, fenômenos lingüisticamente complexos e alvos do projeto, tais como a concordância, demonstraram que a abordagem determinística era insuficiente no tratamento dos desvios gramaticais. Adotou-se, então, uma estratégia mais abrangente, guiada por um *parser*, em que as sentenças do texto de *input* são analisadas em termos de sua estrutura e relações de dependência entre os constituintes. Segundo Bonfante e Nunes (2000), “o *parser* utiliza uma gramática livre de contexto com técnicas de restrição baseadas em relaxamento para permitir o casamento de sentenças agramaticais” (BONFANTE e NUNES, 2000, p. 15). Desse modo, por exemplo, é possível detectar desvios gramaticais no texto do usuário pela reconstituição ou pelo relaxamento da estrutura (ideal) da sentença. Para tanto, o *parser* desencadeia uma série de procedimentos de análise sintática que inclui desde a

classificação sintática dos segmentos sentenciais até o reconhecimento das relações de dependência entre esses segmentos, o que permite, enfim, a detecção de desvios de concordância nominal e verbal, por exemplo.

As regras do *parser* são modeladas em ATNs (*Argumented Transition Networks*), permitindo que as regras pertinentes ao tratamento mecânico, ortográfico e gramatical sejam disparadas imediatamente ao reconhecimento dos *tokens* e detecção de erros. Por sua vez, o módulo lingüístico foi modelado pela filosofia da revisão intercalada (*Interleaved Parsing and Checking*); ou seja, o *parsing* e o *checking* são disparados alternadamente após o reconhecimento de um delimitador de sentença, permitindo que procedimentos de *checking* sejam acionados imediatamente, a partir da validação da estrutura sintática da sentença pelo *parsing*.

Convém mencionar, a respeito dessa arquitetura, que o ReGra conta, hoje, com aproximadamente 1000 regras sintáticas, o que representa um leque bastante amplo de possibilidades no tratamento gramatical. O *parser* é robusto e eficiente, diante das limitações gerais do revisor (como trabalhar em contextos irrestritos), e a inter-relação consistente entre os componentes e subsistemas delineia uma atuação positiva para a aplicação, i.e., efetivamente o ReGra funciona e é atraente ao seu público-alvo.

No entanto, em sua trajetória de desenvolvimento o ReGra permaneceu, ao mesmo tempo, como produto e objeto de pesquisa propriamente dito. Assim, não só os recursos e ferramentas têm sido constantemente ampliados, revisados ou alterados, como as próprias estratégias metodológicas da revisão têm sido reavaliadas.

O ReGra não funciona plenamente, quer em função de desafios na implementação, quer em função da natureza lingüística complexa dos textos que recebe. Suas limitações, portanto, constituem fatos de incômodo que, ao longo do tempo, estimularam diversos trabalhos paralelos (acadêmicos ou não) no próprio NILC.

No caso do *parser*, por exemplo, Bonfante e Nunes (2000) é uma iniciativa de se estudar alternativas para a modelagem descrita atrás. As autoras defendem uma abordagem empírica para o *parser* a qual, por exemplo, resultaria na separação do *parsing* e do *checking* que, como vimos, operam de modo intercalado atualmente. Já em Nunes (2003), constata-se a iniciativa de alteração substancial desse *parser* hoje empregado no ReGra. O *parser* “Curupira” acompanha a concepção de separação proposta por Bonfante e Nunes (2000) e, nesse contexto, propõe uma

análise sintática já no texto de *input* do sistema (e não no *input* do módulo lingüístico, como é hoje). Os segmentos sintáticos seriam, então, rotulados com informações ligadas não só a categorias, mas também à dependência e a papel sintáticos.

O tratamento ortográfico do revisor, do mesmo modo, também foi reavaliado em trabalho específico. Em Pelizzoni (2002) propõe-se um novo comportamento do módulo de correção ortográfica, que passaria a aconselhar a alteração baseado em complexos argumentos e heurísticas. Desse modo, por exemplo, a correção ortográfica do ReGra deixaria de apenas detectar erros óbvios de digitação/ortografia, classificando as sugestões de alteração, para interagir com o perfil do usuário, propondo-lhes alteração em função da “utilidade” da intervenção.

Em síntese, falamos de um sistema de revisão com desenvolvimento em curso ainda hoje. Destaque-se o esforço em adequar o sistema a outras plataformas (como é o caso da versão Regra.Net, lançada em 2003) e em aprimorar ou mesmo construir outros recursos de apoio (a base de conhecimento, p.ex.), bem como funcionalidades complementares (consulta a dicionários mais específicos, à gramática do português, etc.).

II.2.3 Avaliação de SPLNs e a mensuração do desempenho na revisão automática da escrita

Um SPLN, como é o caso de um revisor, requer uma avaliação, dado que visa a uma aplicação. Se se trata de um produto, então, essa avaliação torna-se imprescindível. No campo do PLN, o assunto vem consolidando uma área de estudos chamada, justamente, *Avaliação de Sistemas* que, dentre outros, se ocupa do estabelecimento de parâmetros de avaliação, da composição de recursos para esse fim e da uniformização dos principais avanços no sentido de os tornarem *padrões* e, por isso, recomendáveis à ampla comunidade de desenvolvedores.

A avaliação da qualidade de um sistema envolve diversos temas e traduz suas considerações em termos métricos. Os sistemas sofrem *testes*; os testes revelam sucesso e limitação de uma infinidade de apostas de um projeto. Mas, sobretudo, os testes são necessários quando um “engenho” (daí a obrigatoriedade dos testes no campo da “engenharia”) está em desenvolvimento e uso. E reportar os resultados dos testes em termos de medidas precisas e em

formato padrão (visando ao reuso/à reaplicação) é a etapa final desse processo de metodologia aplicada.

A qualidade de um sistema pode ser medida de diversas maneiras, e estudá-las e defini-las compete ao *processo* de avaliação. Antes do processo, contudo, fala-se em *modelos* de qualidade, e nesse estágio a avaliação lida com seus dois pilares: a motivação e o tipo. Não é possível discutir a avaliação da qualidade, sem estabelecer o fundamento inicial: para quê servirá essa avaliação? Em seguida: qual o formato dessa avaliação; ele é compatível/útil/indicado sob qual ponto de vista?

Os testes, nas aplicações em PLN, foram ganhando espaço e complexidade à medida que os sistemas se multiplicaram e se sofisticaram. Do ponto de vista dos gerentes de projeto e/ou dos projetistas sempre foi primordial testar um programa, ou facetas específicas dele, para assegurar-se de sua eficácia ou para detectar como e onde estão as falhas. Nesse sentido, os testes são mecanismos corriqueiros no desenvolvimento de um programa computacional. Mas, do ponto de vista do usuário, a temática dos testes ocupa espaço há bem pouco tempo, desde o início dos anos 1990.

Testar um sistema, a partir dessa época, significou observá-lo sob uma perspectiva, motivação e referências muito bem delimitadas. As respostas de um teste realizado sob a perspectiva do seu próprio desenvolvedor, por exemplo, vai determinar a busca pela qualidade do sistema, que já havia sido programada no projeto. Sob a perspectiva do usuário, contudo, essas respostas determinarão outros parâmetros de qualidade, que podem, inclusive, reformular antigas idéias concebidas para o sistema. Uma filosofia de avaliação de sistema, portanto, nunca antes foi tão amplamente discutida, conferindo um caráter de “especialidade” a um campo de atuação antes restrito ao laboratório de um projeto.

Em termos de mensuração, os avanços no campo da avaliação consolidaram determinados parâmetros de medida, circunscrevendo o alcance das métricas em sua relação com a perspectiva e a motivação dos testes. No âmbito dos revisores automáticos, por exemplo, medidas comumente empregadas nos testes são as chamadas “cobertura e precisão” (*recall and precision*),

porque se adequam ao instrumento padrão nesse tipo de avaliação, que é o texto escrito¹⁰. Enfim, medir e reportar uma medida, no contexto dos SPLNs, afigurou-se paulatinamente como um trabalho de alto rigor, com complexas variáveis e um impacto profundo na avaliação como um todo.

Outro ingrediente fundamental na avaliação de SPLNs são os instrumentos. Projetar e/ou executar uma avaliação de sistema dependerá severamente da qualidade de seus instrumentos. Numa avaliação sob a perspectiva do usuário, por exemplo, será essencial um bom instrumento de interatividade (os *logging programs*, p.ex.), o que não é o caso de uma avaliação na angular do projetista. Para este, a composição de um corpus representativo de seus usuários (i.e., um “corpus de teste”), por exemplo, pode lhe garantir a recolha de indicativos bastante pertinentes e a um custo menor.

Na metodologia de teste de um revisor, são vários os tipos de teste, variando conforme o objeto, a perspectiva e o interesse da avaliação. Nem sempre o corpus se apresentou como uma alternativa dentre os instrumentos de avaliação, mas é evidente a aceitação cada vez maior que os estudos de corpus vêm apresentando na avaliação do desempenho de um revisor automático, quer do ponto de vista do projetista quer do usuário.

Na avaliação de desempenho, os projetistas podem utilizar a sua própria coleção de dados de referência (até podem ser textos) ou uma lista de sentenças especialmente preparada para o teste, ou um corpus. Denomina-se *corpus de teste* aquele que, especificamente modelado, apóia a tarefa da avaliação no quesito desempenho. Sua representatividade, grau de autenticidade, anotação e documentação será dependente do revisor, de seu usuário, das especificidades, enfim, desse sistema.

Uma avaliação baseada em corpus, em geral, utiliza corpus fechado, i.e., aquele que uma vez concluído não aceita mais inserção de amostras. Quando a avaliação visa ao usuário, porém, é comum encontrarmos mecanismos de interatividade via internet cujo resultado é a compilação de amostras para a composição do corpus de teste. Nesse caso, o modelo de corpus é cíclico e, conseqüentemente, o corpus de teste será do tipo aberto, monitor.

¹⁰ No âmbito da detecção de erros numa avaliação de desempenho de um revisor automático, *cobertura* seria o número de intervenções desejadas pelo total de intervenções numa página, e *precisão*, o número de intervenções desejadas pelo número real de intervenções numa página.

Dada a profunda influência que o corpus de teste tem na avaliação como um todo, é essencial que seu *design* seja modelado minuciosamente. Medidas de precisão serão aplicadas ao corpus e, por exemplo, uniformidade ou balanceamento de gênero e tipo textual são desejáveis na etapa de preparação. A variedade de versões, em certa medida, também traz vantagens para a avaliação; ou seja, quando o corpus de teste compõe-se de textos de referência (versão texto cru) e de textos com desvios de escrita (versão texto anotado), por exemplo, essas duas versões aumentam a capacidade de teste de um corpus.

Podemos notar esse tipo de preocupação num trabalho de avaliação desenvolvido por pesquisadores da Microsoft. Em Helfrich e Musci (2000) reporta-se a iniciativa de avaliação de três revisores gramaticais automáticos (do francês, do alemão e do espanhol) com vistas ao aprimoramento dos sistemas. Ainda que tal avaliação tenha sido abrangente (p.ex., diferentes instrumentos foram utilizados em etapas distintas do processo de avaliação), no estudo em questão o foco é a metodologia de avaliação baseada em corpus. Os autores enfatizam de início: “o papel central de grandes corpus nas três línguas (incluindo variantes regionais), tanto para o design como para a avaliação [dos revisores], será discutido aqui” (HELFRICH e MUSIC, 2000, p. 1036)¹¹.

Em face dessa disposição, reporta-se, então, o perfil do corpus necessário à avaliação. Questões como a extensão de texto e de corpus (representatividade) ou o balanceamento das amostras (optou-se pelo balanceamento de tipo textual, p.ex.) demonstram o cuidado em estabelecer diretrizes minuciosas na preparação do corpus de teste. Destaque-se, no estudo dos autores, a decisão de se trabalhar com textos não-editados, tendo em vista a avaliação do desempenho dos revisores em textos com problemas de inadequação gramatical – justamente aqueles em que a revisão automática seria útil. Soma-se, ainda, a composição do corpus de teste que, para o fim de avaliação sistemática, conta com duas versões: a versão “clara” (*non-blind corpus*), utilizada tanto pelos projetistas dos sistemas quanto pela equipe externa de testes, e a versão “cega” (*blind corpus*), a que apenas a equipe de teste tem acesso.

Nesse tipo de estudo, enfim, observamos que, ao contrário do que encontrávamos há algum tempo, as técnicas estatísticas direcionadas ao teste de desempenho de um revisor não suplantam a necessidade de se modelar criteriosamente o corpus de teste. Em vez disso, tais

¹¹ “The central role of large text corpora in the three languages (including regional variations) for both design and evaluation will be discussed.”

iniciativas parecem envolver-se mais com as proposições da Lingüística de Corpus e o que vemos, então, são trabalhos de avaliação interessados no viés “baseado-em-corpus”, com toda a complexidade que esse comprometimento trará para a avaliação.

E se chegamos a esse ponto, parece-nos que esta exposição deve se aprofundar em assuntos mais afeitos à Lingüística de Corpus.

Nosso leitor pôde notar, até aqui, que a temática da escrita, dos desvios da norma padrão e as especificidades que adquirem no contexto da revisão automática tornam-se cada vez mais complexas. Em termos de pressupostos teóricos da nossa investigação, contudo, há importantes reflexões que devem ser mencionadas aqui. A idéia de uma investigação pautada em observações empíricas (por exemplo, em corpus), tal com vimos enfocando desde a discussão sobre as gramáticas do português, é de extrema importância. Mas, por que, nos estudos da língua escrita, a investigação de um corpus adquiriu importância? Aliás, a noção de corpus foi sempre a mesma nos estudos lingüísticos? O que é um corpus depois do advento do computador? É o que veremos no próximo capítulo.

III. LINGÜÍSTICA DE CORPUS e a fonte de textos do ENEM

III.1 O CONTRAPONTO TEÓRICO

Diversos lingüistas, ao apostarem na indução¹², foram pouco levados em conta no universo chomskiano de prestígio acadêmico. Mas ela ocupou cada vez mais espaço na lingüística à medida que teses como a de Halliday (1991) puderam ser apoiadas pelo avanço tecnológico. A visão chomskiana, segundo a qual o uso de corpus não contribuiria para a investigação sobre as línguas, a não ser como instrumento de corroboração das teses intuitivas que o pesquisador tem sobre a competência do falante (CHOMSKY, 1965 [1975]; HERDAN, 1964), reduzia a credibilidade de uma investigação lingüística de natureza empírica.

Noam Chomsky, nos anos 50 [1950], chamava a atenção para aquilo que considerava ser o verdadeiro exercício do lingüista – o estudo da “gramática universal”¹³. Contrário à idéia de uma lingüística apoiada na observação direta (em geral a partir de um corpus), Chomsky concebe a denominada “lingüística gerativa”, cuja base se constitui de “um sistema de regras que, de um modo explícito e bem definido, atribui descrições estruturais a frases” (CHOMSKY, 1965 [1975], p. 89). Tratava-se, em linhas gerais, de uma teoria lingüística com apelo à descrição e explicação de fenômenos circunscritos à chamada *competência* de um falante/ouvinte-nativo tido como *ideal*, e em cujos princípios metodológicos encontrava-se o emprego de modelos *racionais* do funcionamento cognitivo da linguagem humana.

A crítica de Chomsky aos princípios e métodos da lingüística empírica recai, pois, sobre dois aspectos. O primeiro é o fato de que as duas noções polares *competência* e *performance*¹⁴ devam ser distinguidas, cada qual resultando numa teoria específica, donde a primeira (a teoria da

¹² Segundo o *Dicionário de Lingüística* (Dubois et. al, 1993), o “método indutivo consiste, em lingüística, a recolher um ‘corpus’ e a tirar [dele] listas de elementos e regras”.

¹³ Trata-se do estudo sobre a estrutura e o funcionamento dos “universais lingüísticos”, concebidos pela lingüística gerativa, partindo do princípio de que as línguas têm entre si regularidades, elementos recorrentes e sistematizados, decorrente do fato de que a linguagem possui um componente inato no ser humano. Cf. Chomsky, 1957; 1965.

¹⁴ Neste trabalho, empregamos o termo “performance” (inglês) em vez de “desempenho” (português) ao abordarmos a teoria gerativa. A palavra “desempenho”, para nós, estará restrita à discussão computacional, onde o termo é largamente empregado pela literatura especializada. Não obstante, no português europeu usa-se “competência e performance”, o que nos conforta quanto ao emprego do estrangeirismo aqui.

competência) tenderia à lingüística científica em detrimento da segunda (a teoria da performance). Sendo competência “o conhecimento que o falante-ouvinte possui da sua língua” e performance, “o uso efetivo da língua em situações concretas” (CHOMSKY, 1965 [1975], p. 84), para o autor, o estudo da performance pressupõe a competência. Assume-se, além disso, que uma “teoria da performance” se limitaria a inventariar e descrever os enunciados “aceitáveis” e “inaceitáveis”¹⁵ detectados no uso da língua.

Nessa perspectiva, das críticas de Chomsky salvaram-se, tão somente, as pesquisas empíricas ligadas à fonética, segundo o autor, as únicas que, à época, desenvolviam-se com vistas a extrair elementos regulares e produtivos, i.e., a competência subjacente, na produção da fala. Assim considerados, na visão do autor, os estudos da performance são estritos por questão de princípio e de objetivo científico:

Em geral parece que o estudo de modelos da performance incorporando gramáticas generativas [sic] pode ser frutuoso; mais ainda, é difícil imaginar qualquer outra base sobre a qual uma teoria da performance pudesse ser desenvolvida. Tem havido um certo número de críticas ao trabalho da gramática generativa com base na acusação de descuidar o estudo da performance em favor do estudo da competência subjacente. Os factos, contudo, parecem mostrar que os únicos estudos da performance, fora do domínio da fonética, são aqueles levados a cabo como produto acessório do trabalho em gramática generativa. (...) Consequentemente, esta crítica não se justifica, e, mais ainda, tem um alvo completamente errado. Pelo contrário, aquilo que impede o desenvolvimento duma teoria efectiva da performance é a limitação-de-princípio dos descritivistas à classificação e à organização dos dados, à ‘extracções de padrões’ dum corpus de fala observado, à descrição de ‘hábitos de fala’ ou de ‘estruturas habituais’ (e ainda no caso de tudo isso existir), etc. (CHOMSKY, 1965 [1975], p. 96-97, grifo nosso)

Os críticos de Chomsky estavam corretos ao interpretarem a inclinação do lingüista para a investigação da competência e não da performance e, de resto, nenhuma vantagem levaria

¹⁵ O termo “aceitável”, relativo à performance, está para “gramatical”, no âmbito da competência, assim como “inaceitável” para “agramatical”. Emprega-se a noção de aceitabilidade, na teoria da performance, “para aludir aos enunciados perfeitamente naturais e imediatamente compreensíveis sem uma análise escrita, e que não são de modo nenhum bizarros ou estranhos [segundo o falante/ouvinte]” (CHOMSKY, 1965 [1975], p. 91). Já a noção de gramaticalidade, na teoria da competência (que é a própria lingüística gerativa), faz referência à “boa-fomação” de enunciados; é, em última instância, um “julgamento” que o falante/ouvinte faz sobre os enunciados de sua língua, amparados unicamente na gramática internalizada que, por definição, possui. Há, contudo, “graus” de aceitabilidade e gramaticalidade. Certos enunciados, assim, podem estar num grau baixo da escala de aceitabilidade, mas num alto grau da escala de gramaticalidade. Essas escalas não coincidem, segundo a teoria gerativa, já que “a gramaticalidade é apenas um dos muitos fatores que, pela sua ação recíproca, determinam a aceitabilidade. Por isso mesmo, embora se possam propor vários testes operacionais para a aceitabilidade, é pouco provável que um critério operacional necessário e suficiente possa ser inventado para a noção muito mais abstrata e de longe mais importante de gramaticalidade” (CHOMSKY, 1965 [1975], p. 92, grifo nosso).

Chomsky em proceder ao contrário em vista do modo como ele próprio confinou a teoria da performance a segundo plano.

O segundo fator de descontentamento de Chomsky em relação à abordagem empírica diz respeito à filiação do autor à chamada “lingüística mentalista”. Considerando a distinção entre competência e performance e tendo advogado em favor da relevância de uma sobre a outra, o autor reconhece nos métodos dedutivos e na introspecção o sentido do trabalho do lingüista:

Para o lingüista, assim como para a criança que aprende a língua, o problema consiste em determinar, a partir dos dados da performance, o sistema subjacente de regras que foi dominado pelo falante-ouvinte e que ele põe a uso na performance efectiva. Logo, no sentido técnico, a teoria lingüística é mentalista, na medida em que tem como objetivo descobrir uma realidade mental subjacente ao comportamento efectivo. O uso observado da linguagem ou as disposições possíveis de resposta aos estímulos, hábitos, etc., podem fornecer informações relativas à natureza desta realidade mental, mas certamente não poderão constituir o objectivo efectivo da lingüística, se quisermos que ela seja uma disciplina séria. (CHOMSKY, 1965 [1975], p. 84)

Em nota, o autor acrescenta que “a lingüística mentalista é simplesmente a lingüística teórica que utiliza a performance como um dado (juntamente com outros dados, por exemplo, os dados fornecidos pela introspecção)” (CHOMSKY, 1965 [1975], p. 292). A partir disso, fica fácil derivar que a confecção de corpus, para Chomsky, era serviço inútil, contrário à racionalidade que se impõe à lingüística enquanto ciência lógica.

Implicava também defender que os “dados”, i.e., as evidências lingüísticas observadas a partir de um corpus, devessem ser manipulados para o fim de se compreender os limites da influência da situação de produção na aceitabilidade ou não de um enunciado, e apenas isso. Assim, Chomsky nos leva a crer em relação às frases, por exemplo, que postular sobre a probabilidade de frases aceitáveis ser maior do que a de frases inaceitáveis seria mero recurso de oratória, já que essa probabilidade constitui-se uma noção “vaga, [cujos] conceitos em jogo são igualmente obscuros” (CHOMSKY, 1965 [1975], p. 293).

Por isso, nessa perspectiva mentalista, fala-se em “possibilidade” no lugar de “probabilidade”, em “língua ideal/possível” vs. “língua real/provável”. Enfim, para Chomsky, a noção de *probabilidade*, se tivesse valor em lingüística, teria de ser tratada de modo objetivo, claro e preciso, o que reconhece ser tarefa inalcançável de um lingüista. Considera, pois, o autor

que no âmbito da linguagem humana as interferências do meio (contexto ou situação de produção) sempre serão significativas, sendo necessário recorrer à *probabilidade relativa* dos dados. Esse processo de análise, contudo, não seria confiável:

“na realidade, [a probabilidade de uma frase ou probabilidade de um tipo de frase] serão noções objetivas e claras apenas se a probabilidade for baseada numa estimativa da frequência relativa.” (CHOMSKY, 1965 [1975], p. 293)

A probabilidade relativa, i.e., a frequência de ocorrência de um dado lingüístico em relação ao contexto em que ocorre, não tem credibilidade na abordagem chomskyana, porque esse contexto seria sempre “interpretado” (pelo informante ou pelo próprio lingüista-investigador), segundo o autor, não consistindo, pois, um dado neutro, isento, sobretudo mentalista.

Na abordagem empírica de M.A.K. Halliday, por outro lado, tal questão não poderia ser colocada nesses termos. E será esse posicionamento teórico de Halliday, justamente, que formará o quadro conceitual da Lingüística de Corpus.

Halliday (1991) tensionou, de modo direto, essa posição de Chomsky em relação à probabilidade relativa:

Claramente rejeitada [por Chomsky] foi a relevância de qualquer tipo de dado quantitativo. A observação sarcástica de Chomsky de que ‘Eu moro em Nova Iorque’ é mais freqüente do que ‘Eu moro em Dayton Ohio’ foi lançada para destruir a idéia de que a frequência relativa num texto poderia ter algum significado teórico. (HALLIDAY, 1991, p. 30)¹⁶

Convicto da importância de um corpus para os estudos lingüísticos, Halliday refere-se à probabilidade como algo inerente à língua, enquanto sistema, e afirma que a frequência de um dado lingüístico não pode ser interpretada meramente como acidente ou efeito circunstancial, tal como na teoria gerativa.

Com efeito, os estudos baseados em corpus devem revelar esse caráter probabilístico da língua. Neles, a probabilidade relativa é tão somente uma questão de interpretação dos dados e

¹⁶ “Explicitly rejected was the relevance of any kind of quantitative data. Chomsky’s sarcastic observation that ‘I live in New York’ is more frequent than ‘I live in Dayton Ohio’ was designed to demolish the conception that relative frequency in text might have any theoretical significance.”

não de construto sobre o funcionamento lingüístico. No raciocínio de Halliday abaixo, essa discordância no foco de discussão aparece de modo muito claro:

Parece-me óbvio que, dada a ‘polaridade’ de um sistema cujos termos fossem ‘positivo/negativo’, o fato de o positivo ser mais freqüente do que o negativo seria uma propriedade essencial do sistema – tão essencial quanto os termos da oposição propriamente. Analiticamente, é claro, foi necessário separar as considerações sobre os termos do sistema das considerações sobre as suas probabilidades relativas; mas o que esteve envolvido foi um único fenômeno, não dois. (HALLIDAY, 1991, p. 31)¹⁷

De certo modo, o ponto fundamental na argumentação de Halliday não é o fato de determinada ocorrência ser ou não provável, mas o fato de sabermos com qual freqüência ela acontece. Segundo Berber Sardinha (2004), “a visão da linguagem como sistema probabilístico pressupõe que, embora muitos traços lingüísticos sejam possíveis teoricamente, não ocorrem com a mesma freqüência” (BERBER SARDINHA, 2004, p. 30-31).

Na verdade, o levantamento de freqüências tem história na lingüística. Destaque-se, num período recente, os trabalhos de C. Muller (1960 em diante) a respeito da “estatística léxica” (cf. Muller, 1964, p.ex.). A estatística léxica é uma abordagem baseada em métodos estatísticos com o propósito de descrever o léxico. Nesse campo, estudam-se concepções relativas à amostragem, generalização dos dados e tudo isso, naturalmente, a partir de levantamento de freqüência de um corpus. A lexicologia quantitativa¹⁸ é outro campo para o qual essa idéia em torno da freqüência se tornou um princípio. Contribuição significativa encontra-se na discussão envolvendo freqüência, probabilidade e a relação disso com o uso lingüístico. A lexicografia, por essa perspectiva, passou a refletir, nos dicionários baseados em corpus, a preocupação em associar o sentido de uma palavra ao seu uso, fundamentando-se na idéia de freqüência e *contexto de ocorrência*¹⁹.

Em meio a esse movimento, é a figura de Douglas Biber que oferece para a Lingüística de Corpus um repertório conceitual de importância. Em seus trabalhos, desenvolve-se a noção de

¹⁷ “It seemed to me self-evident that, given a system ‘polarity’ whose terms were ‘positive/negative’, the fact that positive was more frequent than negative was an essential property of the system – as essential as the terms of the opposition itself. Analytically, of course, it was necessary to separate the statement of the terms of the system from the statement of their relative probabilities; but what was involved was a single phenomenon, not two.”

¹⁸ No Brasil, M.T.Biderman (1978) é uma das referências pioneiras.

¹⁹ No caso brasileiro, o *Dicionário de usos do português do Brasil* (Borba et al, 1974) e o *Dicionário UNESP do português contemporâneo* (Neves, 2005) são exemplos de trabalhos lexicográficos com enfoque probabilístico.

“correlação”, por exemplo. Dedicando-se ao léxico, suas idéias em torno da correlação serão relevantes para o delineamento de metodologia quantitativa e influenciarão os princípios do *design* de um corpus. Conforme resume Berber Sardinha (2004) a respeito da contribuição do lingüista norte-americano, conterrâneo de Chomsky:

“[para Biber], há uma correlação entre características lingüísticas e situacionais (os contextos de uso). O conjunto da pesquisa desenvolvida por Biber apresenta evidências inequívocas de que conjuntos de traços lingüísticos variam sistematicamente com relação a textos típicos de contextos comunicativos específicos. Em outras palavras, a variação não é aleatória.” (BERBER SARDINHA, 2004, p. 31)

Tomemos, justamente, esse entendimento sobre a língua: a frequência observada dos dados não acontece de forma aleatória; ou seja, há repetição tanto da ocorrência de um dado lingüístico quanto da situação em que é produzido. Essa noção se tornaria cara à LC, sobretudo porque, a partir disso, tornou-se inevitável demonstrar que a língua é “padronizada”. Destacando-se, em lingüística, os estudos relacionados ao léxico, a questão da *padronização* aparece com força na detecção de elementos, que são de tal maneira regulares e sistemáticos, que se podem chamá-los padrões “léxico-gramaticais” (HALLIDAY, 1991), numa alusão à divisão gramática vs. léxico da teoria gerativa que, aqui, foi dissolvida.

A Lingüística de Corpus, desse modo, instaura-se com a clara convicção de que as línguas são sistemas probabilísticos. De seus primórdios para cá, os avanços das pesquisas envolvidas com corpus parecem indicar que mais que um simples registro da performance do falante, o corpus é um lugar de experimentação dos mais ricos e surpreendentes (FILLMORE, 1992), tanto no que se refere aos resultados com ele obtidos quanto no que diz respeito ao posicionamento teórico que o investigador é levado a assumir ao construir ou utilizar um corpus.

Hoje, é indiscutível o interesse de grande número de lingüistas em basear suas teses sobre a linguagem no conhecimento apresentado pelo mundo real. Esse mundo pode ser, justamente, aquele que se concretiza nos textos escritos e/ou falados, de modo autêntico, em qualquer língua natural – i.e., um corpus (MCENERY e WILSON, 2001). O corpus apresenta-se na lingüística da atualidade não só como meio de extração de informações sobre língua(s), mas também como objeto de pesquisa propriamente dito, devendo o lingüista (de corpus) preocupar-se em conferir-lhe a confiabilidade científica até então reclamada.

Em certo sentido, podemos dizer que um novo paradigma se coloca frente à visão chomskiana sobre a linguagem, indicando que mais (ou melhor) do que descrever o que é *possível* na língua, seria investigar o que nela é *provável* (HALLIDAY, 1991); que antes considerar a *recorrência* (padronização) de dados lingüísticos do que apostar na *ocasionalidade* desses elementos no uso da língua (BIBER et al., 1998; SINCLAIR, 1996).

É preciso registrar, por oportuno, que esse referencial teórico incide de modo positivo na consolidação da LC, mas há fatores externos (ou contextuais) sem os quais a abordagem da LC não teria êxito. Esse componente contextual, de forma unânime na literatura, aparece como sendo “a peça que faltava” no percurso da LC. E é sobre ela que discutiremos a seguir.

III.2 CONTEXTO – A ERA DA INFORMÁTICA

Com uma capacidade sobre-humana de armazenamento de dados, o computador proporcionou ao homem a facilidade de manipulação de grande volume de informações. Sem o computador, porém, a história da LC registra fatos surpreendentes, que nos fazem dimensionar a importância desse invento no cenário da lingüística de que falamos atrás. Kennedy (1998) lembra, por exemplo, que Govindankutty e equipe, nos idos de 1970, chegaram a construir manualmente um corpus com 300.000 palavras com vistas à criação de um Index do Kamparāṁṣānam²⁰. Para isso, contudo, levaram anos para compilar os registros, além de requerer esforço humano gigantesco. Isso revela o impulso que a informática ofereceu ao estabelecimento da LC tal qual a vemos hoje, uma época em que o trabalho de Govindankutty seria efetuado em poucos dias²¹.

O advento do computador representou uma verdadeira revolução no universo dos estudos lingüísticos (ATKINS e ZAMPOLLI, 1994). Como vimos na seção anterior, nos trabalhos ligados ao léxico, por exemplo, o salto que as abordagens ligadas à estatística (computacional) representaram tem poucos equivalentes na história da lingüística. Para as pesquisas com corpus,

²⁰ Índice que associa ocorrências e suas linhas de concordância, com etiquetagem gramatical e semântica. Cf. Govindankutty (1973).

²¹ Técnicas avançadas, p.e., usando XML; capacidade, portabilidade e virtualização do armazenamento de informações, proporcionadas pelos CD-ROMs, internet, servidores cada vez mais potentes: esses são fatores que dão a entender o grau de evolução dos recursos, que também atinge a engenharia de hardware, tão relevante quanto a de software para o sucesso da LC. Além disso, para essa visão devemos considerar a importância do movimento de construção de corpus em larga escala (*large corpora*), especialmente no final dos anos 1990.

então, o computador significou ainda mais: a possibilidade inédita de tratar grande volume de dados num curto espaço de tempo.

A força dessa constatação faz surgir aquilo que se entende, propriamente, por Lingüística de Corpus. Conforme Berber Sardinha (2000; 2004), pode-se constatar significativos avanços no plano teórico em prol da LC, mas esses nada representariam não fosse a criação do computador. O conhecimento tecnológico, estritamente, favoreceu integralmente qualquer iniciativa empreendida, a partir de então, pelos pesquisadores adeptos da análise empírica sobre a linguagem. Como veremos neste trabalho, a própria definição de corpus só tem sentido em função da inserção do computador no universo dos estudos lingüísticos. Antes disso, corpus não era reclamado como objeto de pesquisa de qualquer área do conhecimento, tampouco agregava em torno de si perfis tão variados de pesquisadores, levando-os a se reconhecerem, essencialmente, “lingüistas de corpus”, tal como aconteceu a partir da década de 1990 (a “3^a geração da LC”, conforme McEnery e Wilson, 2001; Leech, 1991).

Para os lingüistas, tarefas antes penosas, como o armazenamento de textos e a contagem de freqüência, por exemplo, tornaram-se automatizadas; a comparação de dados entre línguas distintas ou entre tempos históricos distantes tornou-se possível com bases de dados extensas e analisáveis eletronicamente. Berber Sardinha (2004) resume bem a importância dos computadores no desenvolvimento da LC:

“Nos anos de 1960, os computadores ‘mainframe’ passaram a equipar centros de pesquisa universitários e foram aproveitados para a pesquisa em linguagem. A popularização dos computadores possibilitou o acesso de mais pesquisadores ao processamento de língua natural enquanto a sofisticação do equipamento permitiu a consecução de tarefas mais complexas de forma mais eficiente, já que o aumento da capacidade de armazenamento e a introdução de novas mídias (fitas magnéticas, em vez de cartões ‘hollerith’ perfurados etc.) facilitaram a criação e manipulação de corpora em maior número. Com a entrada em cena dos microcomputadores pessoais, nos anos de 1980, uma nova onda de mudanças aconteceu, como a popularização de corpora e de ferramentas de processamento, o que contribuiu decisivamente para o reaparecimento e fortalecimento da pesquisa lingüística baseada em corpus.” (BERBER SARDINHA, 2004, p. 5)

Essa pesquisa lingüística baseada em corpus, de que fala Berber Sardinha acima, é a que tem efetivamente o aparato eletrônico. Antes dele, a pesquisa lingüística utilizava uma “pseudo-metodologia”, i.e., o corpus manual, para promover análises que, como discutíamos atrás, falhavam no aspecto da confiabilidade. Assim, conforme McEnery e Wilson (2001): “a

tecnologia permitiu que pseudoprocédimentos se tornassem uma metodologia lingüística valiosa”²² (idem, ibidem, p. 18-19). Ou seja: os recursos da informática mostraram-se como fortes aliados da lingüística empírica, sobretudo pela precisão ao lidar com grandes bases de dados. Conseqüentemente, a falta de confiança nos métodos apoiados em corpus passou a ser uma questão de posicionamento teórico e não mais de princípios técnicos.

Se tomarmos alguns dos atributos de um computador, não nos surpreenderemos com tal impacto no campo dos estudos lingüísticos. De acordo com Smith (1991),

Computadores são processadores de símbolos, capazes igualmente de trabalhar com números ou símbolos lingüísticos. A memória do computador pode armazenar grande quantidade de dados de língua e torná-los acessíveis muito rapidamente. Os computadores também podem apoiar nos estudos sobre as propriedades lingüísticas e a cognição envolvendo a língua. Seus processadores podem realizar qualquer regra lingüística que tenha sido explicitada. Até aqui, eles podem reconhecer ou gerar sentenças legítimas de uma língua, produzir parágrafos aceitáveis ou traduções de língua e fazer inferências além de descobrir explicações. O quanto das capacidades humanas que não pode ser emulado com a manipulação de símbolos está sendo reduzido constantemente.²³ (SMITH, 1991, p. 4)

Com efeito, tais funcionalidades de um computador vão ao encontro dos interesses das pesquisas lingüísticas. Segundo McEnery e Wilson (2001), esses interesses resumem-se a quatro processos com dados lingüísticos: busca, recuperação, seleção e cálculo. Tomemos esses processos como ponto de partida a fim de notarmos que, em cada um, determinadas expectativas de avanço tecnológico já foram cumpridas, interferindo substantivamente no campo da LC:

✍ **processo de busca:** as informações estão disponíveis em formato eletrônico e podem ser compartilhadas. Podemos usar esse processo para compilar um corpus. Um corpus, por menor que seja, oferecerá a matéria-prima para o estudioso da linguagem que pode contar, ainda, com mecanismos potentes de mapeamento.

✍ **processo de recuperação:** as informações são extraídas (podendo ser transportadas) e, invariavelmente, o contexto de uso está agregado. Para a pesquisa lingüística, é essencial esse vínculo com o contexto e, a seguir pelas novas tendências, esse processamento se tornará cada vez mais lingüisticamente sofisticado.

²² “Technology has allowed a pseudo-procedure to become a valuable linguistic methodology”.

²³ “Computers are processors of symbols, equally capable of working with numerical or linguistic symbols. Computer memory can store language in enormous quantities and make it available to very rapid searches. Computers can also help in studying properties of language and of language cognition. Their processes can carry out any linguistic rules that can be made explicit. Presently they can recognize or generate legitimate sentences of a language, produce acceptable paraphrases or translations of language, and make inferences and discover explanations. The range of human capabilities that cannot be successfully emulated by manipulating symbols is steadily shrinking.”

✍ **processo de escolha:** as possibilidades de análise das informações são tão variáveis quanto o olhar do linguista. Hoje podemos usar o computador para escolhas (p.ex., investigar um fenômeno sintático específico), empregando-se apenas um programa computacional lingüisticamente preparado. É possível desde a análise de um item lexical à de um gênero discursivo, e para isso conta-se com ferramentas computacionais cada vez mais especializadas (p.ex., os concordanciadores), o que, no mínimo, representa otimização do tempo do investigador.

✍ **processo de cálculo:** o suporte matemático-estatístico inerente ao computador coloca à disposição do pesquisador a capacidade de mapear, contar, medir, enfim, de calcular as informações. De um corpus, extraem-se frequências, distribuição, mede-se correlação e probabilidade, etc. com o refinamento cada vez maior da apresentação desse complexo mundo matemático ao linguista.

É curioso assinalar que essa propriedade do computador de calcular com certa isenção teórica, mas com precisão, aparece como resposta a uma crítica de Chomsky à pesquisa empírica. Em Kennedy (1998), a vantagem do computador para a lingüística, dentre outras, é a de que ele “trouxe para os estudos lingüísticos um alto nível de precisão de medidas, o que é importante em toda ciência” (KENNEDY, 1998, p. 5)²⁴. Por sua vez, Leech (1991) é mais explícito no argumento:

Eis aqui uma medida precisa comparável às medidas de simplicidade que Chomsky em seus primeiros escritos considerou a básica para a avaliação de descrições lingüísticas. A lingüística de corpus não tem mais porquê se sentir tímida quanto a suas credenciais teóricas, da mesma forma que a antiga rejeição chomskyana em relação a dados de corpus não tem mais a mesma força.²⁵ (LEECH, 1991, p. 13-14)

Contudo, se com o computador a LC adquiriu essa consistência teórica, é necessária a consciência de que a relação corpus-computador está condicionada à direção tomada pelo projetista ao desenvolver um corpus ou ao ponto de observação em que se coloca o investigador-usuário. Conforme Kennedy (1998) adverte, o corpus reflete o olhar do pesquisador sobre a linguagem. O computador não pode tudo e é falível; por essa razão, um corpus, se não for criteriosamente especificado e preparado, pode limitar a exploração do potencial do próprio computador. Ou seja: o pesquisador ficará restrito, mesmo com o computador, a executar tarefas apenas elementares e, por conseguinte, a explicitar muito parcialmente o fenômeno lingüístico sob sua observação.

²⁴ “Brought to linguistic studies a high degree of accuracy of measurement which is important in all science”.

²⁵ “Here is a precise measurement comparable to the ‘simplicity measures’ which Chomsky, in his earlier writings, regarded as a basis for the evaluation of linguistic descriptions. Corpus linguistics need no longer feel timid about its theoretical credentials, nor does the earlier Chomskyan rejection of corpus data carry such force.”

Isso posto, parece que não será possível escapar de uma máxima de Leech (1991; 1997), segundo a qual, na LC, há que se distinguir e compreender interdisciplinarmente três elementos: o analista humano, o computador e o corpus. Ao nos referirmos, em tom de prognóstico, aos avanços do campo da informática na intersecção com a lingüística empírica, certamente pendemos a superestimar a máquina e não os demais elementos. A questão é que, talvez, ela seja, de fato, o centro desse novo campo de conhecimento.

III.3 DELINEAMENTO DO CAMPO

Trabalhar com corpus, de certo modo, exige clareza do pesquisador frente a esse objeto. Seu posicionamento técnico bem como a(s) tarefa(s) que deseja cumprir com um corpus serão determinantes seja para a confecção do corpus seja para a análise lingüística. Esse cuidado impõe qualidade e delimita a pesquisa, sobretudo ao se pensar que trabalho com corpus, como esclarece Kennedy (1998), pode ser algo para um projetista, para um teórico, para um especialista-investigador ou para o pesquisador-usuário. É, pois, com esse tema a que denominamos “delineamento de campo” que encerramos o nosso panorama sobre a LC.

Na literatura especializada, é comum depararmos com a questão de a LC ser ou não um campo de conhecimento independente. Mais que uma simples questão de status, essa problemática revela que o amadurecimento do campo já é fato, e que preocupações dessa natureza significam disposição dos lingüistas de corpus para a discussão de seus “modelos” de trabalho com corpus.

A LC tem sido tratada de três maneiras: como uma teoria, como uma metodologia ou como uma abordagem. Aqueles que consideram a LC uma *teoria* exploram seus principais conceitos (a probabilidade, p.ex.) a fim de formular uma explicação satisfatória sobre o comportamento da língua em uso. Para os que consideram a LC uma *metodologia*, o que importa são os procedimentos com o corpus, que lhes permitem gerar resultados plausíveis e contundentes sobre a língua. Por sua vez, aqueles que propõem a LC como uma *abordagem* tratam-na em termos filosóficos, admitindo-a como um caminho sem volta, i.e., estudar a língua a partir de corpus é uma questão de perspectiva, que pode ou não ser guiada por teorias

lingüísticas. Também se exploram métodos os mais variados, com o fim único de observação e sistematização das evidências ali presentes.

Berber Sardinha (2004) faz uma reflexão sobre esse assunto e argumenta, por exemplo, que para se considerar a LC uma metodologia é preciso, antes, definir o que se entende por metodologia. Se ela é tomada como “instrumental”, pode-se falar em sintaxe tradicional vs. sintaxe baseada em corpus, por exemplo: “o que mudaria entre essas vertentes opostas seria o instrumental; os dados, a orientação, os pressupostos teóricos, as implicações dos resultados e tudo o mais permaneceriam (BERBER SARDINHA, 2004, p. 36). Por outro lado, se metodologia é entendida como “um modo típico de aplicar um conjunto de pressupostos de caráter teórico (...) [ela] traz consigo algo mais do que simplesmente o instrumental (idem, ibidem, p. 36). Nesse caso, a LC transcende uma disciplina específica e “exemplifica com precisão a prática empirista, [sendo] o pólo mais distante das abordagens racionalistas” (idem, ibidem, p. 36).

De certo modo, a tendência de lingüistas ilustres da LC é a de considerá-la uma metodologia. Dizemos “de certo modo”, porque uma análise mais cuidadosa nos fará perceber que o viés metodológico, para esse grupo, é guiado por fortes postulados teóricos, o que acaba demarcando o campo, ampliando a noção de metodologia tal como vimos acima, mas, sobretudo, indicando que é possível transitar da metodologia para a teoria sem necessariamente caracterizar a LC como uma coisa ou outra.

No entanto, compreender a LC como metodologia parece ser o caso em Leech²⁶, referência consagrada na área. Para o autor, “a lingüística de corpus não diz respeito a um domínio de estudos, mas, antes, a uma base metodológica a serviço da pesquisa lingüística (LEECH, 1992, p. 105)²⁷. Contudo, o próprio Leech reconhece que o plano da observação seduz de tal maneira o pesquisador que é possível falar em LC “não apenas como a metodologia emergente para os estudos da linguagem, mas também como uma nova iniciativa de pesquisa; na verdade, uma nova filosofia de abordar o assunto”.²⁸ (idem, ibidem, p. 106).

Curiosamente para nós, há na argumentação de Leech sobre esse tema uma comparação entre lingüística de corpus e lingüística computacional. Para o autor, elas podem equivaler-se, já

²⁶ Cf., p.ex., Leech 1991; 1992.

²⁷ “Corpus linguistics refers not to a domain of study, but rather to a methodological basis for pursuing linguistic research”.

²⁸ “Not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject”.

que em ambas tem-se o computador como o instrumento que as define. O computador é o que imprime o tom de metodologia a ambas, muito embora o corpus, ao ser manipulado “por meio” da máquina, associe um componente especial à LC, que é *revelar a língua em uso*, o que não acontece com a lingüística computacional em relação aos seus objetos de construto.

Essa noção, digamos, “iluminativa” da LC tem sido enfatizada inclusive por Ingüistas que são caros à lingüística computacional. C.J. Fillmore, por exemplo, num texto seminal para a LC, admite que seu interesse por corpus baseia-se unicamente no fato de que as evidências ali reveladas não poderiam, na maioria das vezes, ser detectadas a priori. A certa altura do texto, o autor não se contém em declarar:

“Tenho duas observações a fazer. A primeira é que não acho que possa haver qualquer corpus, por maior que seja, que contenha informações sobre todas as áreas lexicais e gramaticais do inglês que eu quisesse explorar; tudo que tenho visto é inadequado. A segunda observação é que todo corpus que tenho tido a chance de examinar, ainda que pequeno, tem me revelado fatos que eu não imaginaria encontrar de outro modo”. (FILLMORE, 1992, p. 35)²⁹

A tomar pelo prestígio que Fillmore tem na lingüística contemporânea (dentre elas, a computacional), é possível dizer que boa parte dos lingüistas tem recuado nas críticas à LC, e até tomado o seu partido, afinados à idéia de que o corpus mostra aquilo que não saberíamos ver de outro modo.

Tal tendência iluminativa da LC, contudo, propõe determinadas estratégias de trabalho consideradas insatisfatórias, porque dependentes de teorias e limitadoras das potencialidades de um corpus. Argumenta-se, por exemplo, que na descrição lingüística manipulam-se noções e categorias “cristalizadas” (p.ex.: verbo, sujeito, agente da passiva), numa “rotulação” sobre o funcionamento lingüístico que expressa muito mais a teoria subjacente do que o corpus propriamente dito.

Ora, os lingüistas “de corpus” não tratariam a LC dessa maneira. Esses, chamados por Berber Sardinha, de “autodefinidos”, pensam a LC como um viés, uma “maneira de chegar à linguagem” (BERBER SARDINHA, 2004, p. 37). A LC é tomada como uma abordagem (*approach*), e nela não é possível a co-existência de perfis diferentes de investigação num único

²⁹ “I have two main observations to make. The first is that I don’t think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I’ve had a chance to examine, however small, has taught me facts that I couldn’t imagine finding out about in any other way.”

pesquisador, i.e., aquele que se propõe ao estudo empírico irrestritamente e aquele que limita aquilo que esse empírico tem a revelar. Desse modo, no tocante à questão da precedência e preexistência da teoria ao corpus, esses lingüistas de corpus priorizam o corpus sobre a análise lingüística, o que, no extremo, os leva a dispensar os rótulos dados pela lingüística tradicional, propondo uma análise talvez menos imediatista do corpus em favor da chamada “autodefinição” dos “modelos” de língua (cf., adiante, a visão do estudo “direcionado-pelo-corpus”).

Essa linha de pesquisadores toma a LC em duas perspectivas: a abordagem *baseada-em-corpus* (*corpus-based approach*) e a abordagem *direcionada-pelo-corpus* (*corpus-driven approach*). No debate mais contemporâneo, a idéia de a LC ser uma abordagem significa um novo patamar de sua filosofia. A fim de apresentar alguns pontos importantes dessa discussão, abordamos essas duas perspectivas separadamente.

III.3.1 Abordagem Baseada-em-Corpus

Como vimos até aqui, uma pesquisa com corpus, inevitavelmente, dedica-se à língua em uso. Caracteriza-se, porém, como pesquisa *baseada-em-corpus* aquela que toma o corpus como um meio de atribuir “autoridade” ao trabalho. Quer o interesse seja teórico, quer metodológico, importa notar que, aqui, as evidências de um corpus (ou simplesmente “os dados”) são detectadas (e isoladas, veremos depois), sendo manipuladas com o fim de servir a uma hipótese previamente elaborada para o funcionamento lingüístico. O corpus, assim, corrobora idéias, ou as amplia, no caso de os dados surpreenderem o modelo lingüístico do pesquisador.

Para Tognini-Bonelli (2001), se um trabalho desse tipo toma a LC como sua metodologia, ele “se beneficia do corpus, essencialmente, ao tomá-lo para esclarecer, testar ou ilustrar teorias e descrições formuladas antes que um corpus de larga escala estivesse disponível para subsidiar estudos lingüísticos” (TOGNINI-BONELLI, 2001, p. 65, grifo nosso)³⁰. Se, por outro lado, o trabalho é de enfoque teórico, o trabalho com corpus interessa, porque é conveniente, i.e., o corpus entra com os dados e o pesquisador, com suas hipóteses; o processo é econômico, já que os ajustes da teoria serão pontuais, no caso de o dado surpreender as hipóteses. Nesse caso, prepondera o desejo, segundo a autora, de “validar” determinado aspecto de uma teoria e, para

³⁰ “(...) that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study”.

tanto, as evidências do corpus funcionam como “um bônus extra em vez de um fator determinante da análise, esta ainda conformada a categorias pré-existentes” (idem, ibidem, p. 66, grifo nosso)³¹.

Nas pesquisas baseadas-em-corpus, há uma preocupação irrefutável com a supremacia de dois aspectos ligados aos dados. O primeiro é a quantificação: os dados são passíveis de serem contados e medidos. O segundo é a padronização: a combinação dos elementos lingüísticos não ocorre de forma aleatória; antes, ela é recorrente e regular, ou seja, padronizada. Juntas, a padronização e a quantificação constituem o núcleo duro de um trabalho baseado-em-corpus.

No que se refere à padronização, Biber et al. (1998) introduzem a noção dos “padrões de associação” e, com isso, apresenta a LC-corpus-based como uma perspectiva de descrição lingüística inovadora. Para os autores, “os padrões de associação representam relações quantitativas, medindo a extensão em que traços ou variantes estão associados a fatores contextuais” (BIBER et al., 1998, p. 4)³². Por conta disso, prosseguem os autores, “o objetivo das investigações baseadas-em-corpus não é meramente relatar as descobertas quantitativas, mas explorar a importância dessas descobertas para aprender sobre os padrões de uso da língua” (idem, ibidem, p. 5)³³.

Na verdade, Biber mostra-se modesto na declaração acima, ao conferir secundaridade às informações quantitativas extraídas de um corpus, já que o autor é um dos muitos lingüistas de corpus que se inspiram na caracterização quantitativa dos dados para revelar padrões de língua. Basta tomarmos que o traço quantitativo é tão fundamental à definição dos padrões, que Biber passa a denominá-los “padrões quantitativos”, os quais, segundo o autor, seriam essenciais “ao se propor interpretações funcionais para a explicação do porquê de os padrões existirem” (BIBER et. al, 1998, p. 9)³⁴.

Ao mesmo tempo, a proposta dos estudos baseados-em-corpus contribui, de modo geral, para o amadurecimento tanto dos princípios quanto das técnicas e dos produtos fixados na LC.

³¹ “(...) an extra bonus rather than as a determining factor with respect to the analysis, which is still carried out according to pre-existing categories”.

³² “(...) association patterns represent quantitative relations, measuring the extent to which features and variants are associated with contextual factors”.

³³ “(...) the goal of corpus-based investigations is not simply to report quantitative findings, but to explore the importance of these findings for learning about the patterns of language use”.

³⁴ “(...) to propose functional interpretations explaining why the patterns exist”.

Conforme buscamos demonstrar, torna-se irrelevante rotular o campo como essencialmente metodológico ou teórico; em vez disso, podemos chamá-los simplesmente, uma “abordagem”.

Contudo, no que diz respeito aos dados, justamente, tornam-se característicos da abordagem baseada-em-corpus: o isolamento, a simplificação e a padronização.

De fato, conforme Tognini-Bonelli (2001), pesa o fato de que há nesse tipo de trabalho uma tendência primordial em selecionar um fenômeno, um elemento lingüístico ou um padrão associativo para, a partir daí, dar início a uma análise. Esta, invariavelmente, terá como resultado uma explicação ou exemplificação sempre parcial da língua. Segundo a autora, isso se deve ao isolamento dos dados (que é realizado com certo sucesso) e à limitação deles em relação ao comportamento lingüístico do sistema como um todo.

Por sua vez, ao quantificar e padronizar excessivamente as informações de um corpus, a análise torna cada vez mais cristalizados determinados fundamentos teóricos e categorias lingüísticas definidos a priori. Esse processo simplifica o potencial do corpus e nivela as mais distintas variáveis a um comportamento padrão, ainda que esse padrão somente exista em função de um corpus específico e às vezes exaustivamente preparado para o estudo.

A despeito desse tipo de crítica, o trabalho baseado-em-corpus avança a passos largos e, seguramente, tem contribuído para a maior penetração dos estudos de corpus em diversas áreas do conhecimento, inclusive na própria lingüística.

III.3.2 Abordagem Direcionada -pelo-Corpus

A caracterização da perspectiva *corpus-driven* pode ter como ponto de partida a questão sobre o que se considera “evidência”. Por exemplo: o que vem a ser “evidência” para a visão baseada-em-corpus em contraponto à visão direcionada-pelo-corpus?

Para ambas, é necessário assumir, a língua está em uso e um corpus contém “evidências” desse uso lingüístico. Mas, se tomamos a perspectiva baseada-em-corpus, certamente tenderemos a “informar” expectativas de padrões, revelando o olhar do pesquisador sobre o corpus, talvez de uma maneira preponderante à revelação das evidências ali contidas. Por sua vez, na perspectiva direcionada-pelo-corpus, à luz de Tognini-Bonelli (2001), o que importa saber é “evidência de

quê o corpus nos dá”³⁵ (TOGNINI-BONELLI, 2001, p. 78, grifos da autora), neutralizando a influência do olhar do investigador. Em outras palavras: para ambos os tipos de abordagem, há consenso sobre o fato de que um corpus constitui-se “evidência”; o que difere é: esse corpus é evidência de quê; daquilo que queremos que ele seja ou daquilo que ele é, de fato?

Num trabalho direcionado-pelo-corpus, insiste-se na concepção de que é o próprio corpus que indicará o caminho (teórico, metodológico ou ambos) a percorrer. O corpus é, efetivamente, o foco da atenção, e não a teoria, as hipóteses. Rejeitam-se, então, categorias pré-definidas; prioriza-se a minimização das interferências humanas sobre o texto, de modo que o processamento auxiliado pelo computador (processamento automático) possa revelar atributos lingüísticos, padrões associativos, regularidades, recorrências significativas. Tognini-Bonelli (2001) define da seguinte maneira a abordagem direcionada-pelo-corpus:

“(...) ela almeja derivar sistematicamente categorias lingüísticas a partir de padrões recorrentes e das distribuições de frequência, que emergem da língua em contexto.” (TOGNINI-BONELLI, 2001, p. 87, grifo nosso)³⁶

Nessa perspectiva, promove-se uma clara distinção entre introspecção e observação, já que o conhecimento dos fenômenos lingüísticos decorre da valorização das evidências do corpus tais como elas estão, ie., “em contexto”. No plano teórico, isso significa dizer que a pesquisa direcionada-pelo-corpus cria uma hipótese, se e apenas se, esta refletir a evidência (SINCLAIR, 1991). No plano metodológico, representa dizer que há passos claros e definidos: primeiro, observa-se; depois, constrói-se uma hipótese; em seguida, e por meio da comparação e recorrência, busca-se a generalização e, se tudo isso foi possível, aponta-se uma unificação da concepção teórica (TOGNINI-BONELLI, 2001; SINCLAIR, 1991).

Tal como Sinclair (1992) destaca:

“A língua é um objeto de estudo por si só, e há uma porção de questões de que se ocupar que não tem de ser justificada pela existência de clientes ansiosos.” (SINCLAIR, 1992, p. 380)³⁷

³⁵ “what the corpus gives evidence of”.

³⁶ “it aims to derive linguistic categories systematically from the recurrent patterns and the frequency distributions that emerge from language in context.”

³⁷ “Language is an object of study for its own sake, and there are plenty of enquiries to be undertaken which do not have to be justified by the existence of hungry clients.”

A grande conquista para a LC, na perspectiva desse grupo, fixa-se no fato de que, agora, tem-se *corpus*, ao contrário do que acontecia no início do último século. E, nesse sentido, conforme Sinclair acima, é a existência do corpus (e não a de pesquisadores com suas indagações) o que constitui a novidade da LC. Cabe aos pesquisadores, dentre outros, criar mecanismos lingüístico-computacionais para favorecer a “auto-organização” (*self-organization* – Leech, 1991) dos dados do corpus, a fim de que seja possível extrair dele descrições lingüísticas ricas e robustas.

Por sua vez, é de certo modo frustrante a empreitada desses lingüistas de corpus, que têm sido levados a aprender a se despojar das categorias e até mesmo métodos de pesquisa há muito empregados sem que algo (padrão) exista para colocar em troca. Em Tognini-Bonelli (2001), a autora observa que esse caráter de despojamento é necessário à pesquisa direcionada-pelo-corpus. Trata-se de uma “decisão” de investigação, nem sempre com apelo junto a projetos financiados de pesquisa, que preferem apostar em “dados baseados em modelos”, significando que o corpus deva ser observado “a partir” da intuição do especialista (veja-se o termo “cliente” empregado por Sinclair acima). O despojamento do lingüista direcionado-pelo-corpus refere-se ao abandono dessa “intuição” como ponto de partida da investigação.

Mas, para exemplificar o movimento diferenciado das visões baseada-em-corpus e direcionada-pelo-corpus, vamos retomar o tema da padronização e introduzir o da anotação de corpus. É preciso esclarecer, a essa altura, que a padronização, perseguida no trabalho baseado-em-corpus, também é um caminho e um princípio assumidos na abordagem direcionada-pelo-corpus. Contudo, uma das particularidades do tratamento dos padrões, na perspectiva desta última, é apostar na “objetividade” dos dados e investir fortemente no aprimoramento e inovação de recursos lingüístico-computacionais, bem como de sua base técnica que é a estatística. Assim, prefere-se a dedicação ao texto discreto, chamado *texto cru*, em vez da dedicação ao *texto anotado*³⁸.

Mas, na perspectiva baseada-em-corpus, num texto anotado a “informação adicional” favorece a extração de padrões, a recuperação de dados, a integração com modelos formais de

³⁸ Grosso modo, podemos chamar de “texto cru” aquele que registra apenas o que foi produzido pelo autor. “Texto anotado”, por sua vez, é aquele que carrega informações adicionais, introduzidas pelo lingüista-pesquisador, agregando dados dos mais diversos tipos ao texto original. São exemplos de anotação: morfológica (ex.: pedra = nome feminino); sintática (ex.: em “O menino não viu a pedra”, pedra = complemento verbal/objeto direto); semântico-ontológica (ex.: pedra = ser inanimado e concreto); etc.

processamento lingüístico. Investe-se, pois, na construção dos rótulos (se já não existirem) e no acréscimo cada vez mais ousado dessas informações ao texto cru. A questão é: esses dados adicionais não são, absolutamente, dados objetivos. A “subjetividade” inerente à construção de etiquetas é, por si só, de natureza teórica e ideológica. Por isso, qualquer marca, na anotação lingüística, carrega um significado intencional.

Contudo, os avanços nesse campo da anotação são inquestionáveis, embora essencialmente respondam a uma demanda em alta (ainda) nos dias de hoje. Torna-se relevante destacar, por oportuno, que as duas perspectivas – baseada e direcionada a corpus – podem convergir quanto ao aproveitamento de determinados conceitos (como o da padronização, p.ex.), divergindo, porém, no modo como os qualificam, como os empregam metodicamente e como os sustentam na inovação de recursos tecnológicos ligados a corpus.

III.3.3 Trabalho com corpus: uma questão de posicionamento

As diferentes facetas do trabalho com corpus, tal como vimos, pode levar um iniciante, por exemplo, a sérias dúvidas sobre as bases de suas diretrizes de trabalho ou pior: à instabilidade de suas convicções sobre a confiabilidade quer da construção do corpus quer dos próprios resultados da análise de corpus. É o próprio processo de consolidação da LC (com a busca pela afinação entre as expectativas lingüístico-teóricas e a evolução tecnológica), aliado ao interesse multidisciplinar dos pesquisadores, que torna muitas vezes dispersiva a maneira de abordar e lidar com as várias etapas de um trabalho com corpus.

Iniciamos este tópico da exposição chamando a atenção para a necessidade de o pesquisador ter absoluta clareza sobre dois pontos fundamentais: o primeiro é o repertório técnico da LC que lhe seja útil e compatível, e o segundo é o objetivo que busca atingir através dos estudos em LC. Ao traçarmos um breve panorama do campo da LC, procuramos destacar as diferentes perspectivas que concorrem, hoje, para a consecução de um trabalho com corpus dentro dos parâmetros da LC. Podemos dizer, então, que um primeiro passo para situar uma iniciativa de pesquisa com corpus na LC é o pesquisador instalar-se em um dos tipos de abordagem (baseada ou direcionada a corpus) e definir o seu posicionamento em termos de um pesquisador que trata o corpus com interesses conceituais ou procedimentais (metodológicos).

Contudo, a transparência desse tipo de posicionamento talvez não bastará para o sucesso do trabalho com corpus, e isso nós também acenamos no início desta discussão. À luz de Kennedy (1998), dissemos que será fundamental que o pesquisador delimite o seu ângulo de visão e estabeleça, de uma vez por todas, qual o aspecto (ou a faceta) do trabalho para o qual estará disposto a contribuir e avançar. Falamos, pois, do investigador que se posiciona como o teórico da LC, ou como o usuário/analista de corpus, ou como o projetista de corpus, ou mesmo o especialista computacional, dedicado à modelagem e/ou implementação de recursos de apoio.

Ao se colocar, por exemplo, como usuário de corpus, o investigador terá de lidar com questões e de tomar decisões que são de natureza distinta daquelas de que se ocupará o informata de corpus. Para o primeiro, será decisivo o perfil do corpus de que dispõe bem como o instrumental de manipulação; mas é um fato, desde o início, que o trabalho partirá de um corpus já pronto e, em geral, um corpus refinado, robusto, compatível, enfim, com as expectativas da análise lingüística. Decisão essencial para este, ademais, será a hipótese teórica ou a expectativa de análise construída com vistas ao corpus. Já para o segundo, a preocupação latente é a de “enfrentar” o corpus com o máximo de aproveitamento da tecnologia de que se dispõe no momento. Nesse caso, um único corpus não bastará para comprovar a eficácia dos recursos computacionais em construção, já que essas ferramentas, para serem consistentes, terão de “rodar” em corpus fora da esfera dos testes. E mais: esse informata terá de lidar com a própria questão da expectativa do usuário, a fim de que suas ferramentas de fato contribuam (e façam diferença) para a tarefa final do trabalho com corpus, que é a análise lingüística.

É justo esse tipo de percepção que nos guiará, a partir daqui, na empreitada de construirmos um corpus com o fim específico de investigarmos os desvios da norma padrão, ou seja, na criação do Corvo.

III.4 Corpus de redações do ENEM: um corpus de estudo sobre os aprendizes da língua materna

Nesta seção, para encerrarmos os apontamentos sobre corpus, introduzimos o nosso próprio corpus de pesquisa, mais especificamente a sua fonte de textos – as redações do ENEM. Traremos à discussão, de pronto, o material que adquirimos junto ao INEP e, a partir dele,

buscaremos mostrar algumas facetas do tal posicionamento que o pesquisador de corpus é levado a assumir, especialmente se ele é o construtor e, ao mesmo tempo, o analista do corpus.

As amostras do nosso corpus são extraídas de uma única fonte: o banco de redações do ENEM, ano de 2002. Do acordo firmado com o Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) foi possível obter: a) um Banco de dados: conjunto de informações sobre o perfil dos redatores, relacionado às redações, em formato .sav (planilha estatística); e b) um Banco de redações: conjunto de textos em formato .tif (documento de imagem).

O banco de dados se compõe de quatro itens do perfil dos redatores e foram extraídos de um *questionário sócio-econômico* distribuído aos candidatos. Esse questionário tem 73 questões abordando o contexto social, econômico, cultural e educacional dos candidatos. Desse conjunto, selecionamos apenas os itens que nos pareceram suficientes para “informar” sobre o redator dos textos. Além dos dados do perfil, esse banco contém as notas atribuídas aos textos em cinco quesitos. A configuração dessa base de dados apresentou-se, então, da seguinte maneira:

Tabela III-1: Composição do Banco de Dados do ENEM: material fonte do corpus Corvo

Perfil	Item	Qtde. Variáveis
	<i>Idade</i>	3
	<i>Origem</i>	
	<i>Região</i>	5
	<i>Estado</i>	27
	<i>Instituição</i>	4
	<i>Renda</i>	4
	<i>Nota</i>	
	<i>Global</i>	1
	<i>Competências</i>	5
Total	5	49

A planilha do INEP ainda informa o tipo de participação, distinguindo os inscritos ausentes e presentes à prova de redação. Esse é um dado importante, mas pouco significativo para o nosso projeto e, diante disso, não aparece na tabela acima. Consideremos, então, algumas particularidades desses itens:

☞ *idade*: distingue os candidatos concluintes dos egressos do ensino médio. Os concluintes são os indivíduos com até 18 anos.

☞ *origem*: é uma das duas variáveis sub-especificadas e também a mais extensa, contemplando os 27 estados do país, além do Distrito Federal. Essa variável é também a única que apresenta dados

completos no banco de dados, i.e., todo item do banco possui um “valor”; a variável é exaustivamente preenchida. Isso será de grande importância adiante.

☞ instituição: distingue os candidatos que cursaram o ensino médio em escola pública ou privada, ou alternadamente um e outro.

☞ renda: cumpre ressaltar o valor do salário mínimo (sm) em 2002, que era de \$200,00 (duzentos reais). Em termos gerais, os intervalos de até 5sm e acima de 5sm podem discriminar uma população de renda baixa, média e alta. Por isso, consideramos a variante “nenhuma renda” como um indicador do indivíduo que não trabalhava na época do exame, “até 5sm”, para indivíduos de renda média e “acima de 5sm” para os de renda alta³⁹.

Quanto às notas, as redações foram avaliadas segundo cinco diferentes critérios, chamados “competências” no ENEM. É bem verdade que se trata de um atributo do texto e não do redator, mas, para a construção do nosso corpus, os valores relacionados à nota e suas variantes são tomados como dados objetivos, tanto quanto os itens do perfil.

De acordo com o documento oficial do ENEM (BRASIL, 2002), as cinco competências representam, no escopo da avaliação da redação:

☞ Competência 1 (COMP1): o domínio da norma culta padrão;

☞ Competência 2 (COMP2): a adequação ao tema e ao gênero textual;

☞ Competência 3 (COMP3): a seleção dos argumentos;

☞ Competência 4 (COMP4): a construção da argumentação;

☞ Competência 5 (COMP5): a proposta de intervenção.

☞ Nota global (GLOB): é atribuída pelo exame em função de cálculo específico sobre as cinco competências.

Os valores aparecem, no banco de dados, num intervalo de nota que varia entre >30 e 100, qualquer que seja a variante. Nós separamos essas notas em *faixas de rendimento*, organizando os valores da seguinte maneira:

Tabela III-2: Faixa de rendimento das notas nas redações do ENEM

Faixa	Código	Intervalo de notas
Rendimento Insuficiente	Ri	>30
Rendimento Baixo	Rb	30-49
Rendimento Médio	Rm	50-69

³⁹ Essa divisão não tem critérios técnicos econômicos.

Rendimento Expressivo	Re	70-99
Superação	S	100

O banco de dados do ENEM foi processado com vistas ao corpus de estudo especialmente; ou seja, aplicamos essa classificação do rendimento para separar os textos que interessavam ao corpus. Esse tratamento da base foi importante para determinarmos:

- a exclusão dos casos em que o indivíduo fez o exame, mas a redação foi anulada. Esses casos não acusam valores na variável nota;
- a distribuição e a frequência dos valores nulos (ou campos vazios) em todas as variantes do perfil.

O resultado desse processamento expressa não apenas o status geral da base, como também o volume de dados em questão, e se apresenta conforme a tabela abaixo:

Tabela III-3: Quadro geral do banco de dados do ENEM

Especificação	Frequência	% sob (A)
Total de Inscritos	1.820.170	
Total de Comparecimento no Exame	1.250.834	68,7%
Candidatos que fizeram a redação	1.223.460	97,8%
Redações anuladas	27.374	2,2%
Itens com “pelo menos” um campo vazio	95.541	7,8%
Campos vazios – idade	71.508	5,8%
Campos vazios – renda	80.344	6,6%
Campos vazios – instituição	79.635	6,5%
Itens com “nenhum” campo vazio	1.155.293	94,4%

Daqui por diante, chamaremos esse material de *Base-INEP*. Tal como se observa na tabela acima, na Base-INEP: a) temos um total de **1.250.834** candidatos avaliados pelo exame, dos quais 1.223.460 (97,8%) têm potencial para compor o nosso corpus; mas b) do total de textos, 94,4% oferecem condições reais de mapeamento completo, uma vez que não apresenta campo vazio em qualquer variável do perfil.

A partir dessa tomada de decisão, processamos novamente o banco de dados, buscando preservar: a) somente os casos em que o texto foi avaliado; e b) somente os casos em que todos os campos das variáveis foram preenchidos. Então, denominamos *Base-ENEM* a base resultante desse novo processamento, também podendo ser chamada de “base limpa”.

A Base-ENEM é aquela em que nos basearemos de modo mais imediato na construção do corpus Corvo. Ela se apresenta conforme o quadro a seguir:

Tabela III-4: Quadro geral do Base-ENEM

Tipo de Participação	Frequência	%
Texto não válido	119.406	9,5
Texto válido	1.131.428	90,5
Total de Comparecimento no Exame	1.250.834	100,0

Texto não válido = redações anuladas ou com falta de dados das variáveis

Antes de prosseguirmos com a exposição do processo de delimitação do corpus, é pertinente mencionarmos um dado relevante sobre as várias edições do ENEM.

Em 2001, instituiu-se a isenção da taxa de inscrição no ENEM para o candidato conculinte proveniente da escola pública. Nesse mesmo ano, o salto da participação geral no exame foi gigantesco: de cerca de 220 mil, em 2000, para cerca de 1 milhão e 200 mil em 2001. Considerando-se apenas os concluintes, essa abrangência da participação alcançou a marca dos 47% em 2001 (passando a 50% em 2002); ou seja: do total de alunos concluindo o ensino médio, em 2001, o ENEM atraiu 47%, em 2001, contra cerca de 12% em 2000⁴⁰.

Com efeito, essa isenção da taxa de inscrição, conforme atestam os especialistas da área e o próprio documento oficial do ENEM, exerceu forte influência para a composição do cenário que vimos atrás na Base-ENEM. Assim, parece ser possível comprovar hipóteses político-educacionais com a tabulação dos dados que trazemos com nossa investigação, o que, no entanto, não está no escopo deste trabalho.

Dito isso, voltemos à Base-ENEM e os recortes que a nossa investigação impôs a ela.

O acesso às notas das redações do ENEM permitiu-nos compor o corpus focados nesse parâmetro objetivo. Isso mostrou-se satisfatório, pois estávamos atentos à idéia de reduzir o grau de interferência do projetista na seleção das amostras do corpus. Mas ainda parecia insuficiente a configuração da base de dados para a delimitação do nosso corpus, e prosseguimos, então, com a análise sobre os dados da planilha.

Inicialmente, consideremos a figura a seguir:

⁴⁰ Fontes: BRASIL, 2002; Site do INEP: www.inep.gov.br

Figura III-1: Participação das Faixas de Rendimento em cada Competência na Base-ENEM

Ri-S = faixas de rendimento – Cf. Tab. IV-2

Conforme ficou demonstrado anteriormente, o desempenho geral dos redatores concentrou-se nas faixas Rm e Re. Mais de 80% dos candidatos incluem-se nesse intervalo (cf. Fig. IV-6). Se considerarmos, agora, a distribuição do rendimento pelas cinco competências, vemos que o cenário da nota global se confirma (cf. figura acima). Ou seja, para mais de 80% dos casos o rendimento na redação foi avaliado como médio e expressivo em qualquer competência, recebendo notas que variam entre 50 e 99 pontos (de um total de 100).

Considerando o compromisso do nosso corpus com a investigação do uso da norma padrão pelos redatores do ENEM, julgamos essencial partir da hipótese de que os textos mais fracos na avaliação do ENEM seriam, justamente, os de interesse do corpus. Entendemos que nesses textos a probabilidade de incidência de desvios lingüístico-normativos é maior do que nos textos bem avaliados no exame. Assim, pela necessidade de recortes no material de pesquisa, tomamos esse fator como um ponto de partida no estabelecimento de critérios de seleção textual.

Os textos com nota máxima no exame (aqueles incluídos na faixa “S” – superação) são supostamente livres de desvios lingüístico-normativos. Desse modo, determinamos a exclusão dos componentes qualificados como “S” na variável nota global. A exclusão dessa faixa de rendimento representou um corte de aproximadamente 27 mil textos, ou 2% do total da Base-ENEM.

A análise do quadro III-1 acima apontou-nos, em seguida, que a nota global reflete de forma heterogênea o desempenho do redator nas diferentes competências. Isto é, a nota global

reúne num mesmo grupo, por exemplo, aqueles que foram muito bem avaliados numa competência específica, mas muito mal avaliados em outra. Em outras palavras, notamos que era possível que a nota global com alto valor fosse estabelecida em função de muito bom desempenho em apenas uma ou duas competências.

Sob o ponto de vista da correção lingüístico-normativa, foram os dados sobre a competência 1 que se mostraram, então, significativos. Notamos que o bom rendimento global dos textos era sobrelevado não só pelo desempenho homogêneo dos casos de superação (i.e., se o texto tem nota máxima, invariavelmente é devido ao excelente rendimento em todas as competências), mas sobretudo porque havia uma dentre as competências que impulsionava essa nota global – justamente, a nota da competência 1.

Segundo o quadro III-1, a competência 1 (COMP1) tem a maior progressão nas faixas de rendimento. Em rendimento insuficiente (Ri), COMP1 já se mostra como o critério mais bem aproveitado pelos redatores, já que apenas 3,09% têm nota insatisfatória (o menor índice dentre as competências). Quando observamos o rendimento baixo (Rb), notamos que o distanciamento entre COMP1 e as demais competências se mantém. Em rendimento médio (Rm), COMP1 se equilibra aos valores das demais competências, mas, ainda assim, apresenta um índice considerável: metade dos textos (50,82%) está nessa faixa de rendimento. É observando o rendimento expressivo (Re), porém, que a diferença entre as competências se revela. COMP1 concentra, nessa faixa, um conjunto significativo de textos (37,55%), o que não se verifica nas demais competências, cuja concentração fica abaixo dos 17% (exceto COMP4, com 21,28%). Os casos de superação (S), por fim, confirmam essa tendência positiva de COMP1 ao revelar que maior número de textos foram bem avaliados em COMP1 em relação às outras competências.

A competência 1 será discutida em detalhes adiante. Por hora, importa destacar que essa análise do comportamento das notas entre as cinco competências nos autoriza a definir um novo recorte da Base-ENEM.

Primeiramente, consideramos que o princípio que norteou a exclusão dos casos de superação também se aplica ao rendimento expressivo; ou seja: se é provável que nos textos de superação encontramos baixo volume e variedade de desvios lingüístico-normativos, é

igualmente provável que nos textos de rendimento expressivo esse fato se repita, dada a boa avaliação em COMP1.

Com isso, promovemos novo recorte na Base-ENEM focados na nota global. Contabilizamos, então, uma baixa de aproximadamente 180 mil textos e, com essa configuração, o banco de textos, daí por diante, passou a se chamar “Base-ENEM congelada”.

Até aqui esse processo de delimitação do corpus revelou a importância de uma premissa: a de que é menor a probabilidade de desvios linguísticos ocorrerem em textos bem avaliados em termos gerais (nota global) e em termos específicos (nota da competência 1) do que em textos mal avaliados nesses quesitos. Mas a configuração da Base-ENEM congelada não nos satisfaz plenamente.

Interessados, pois, na norma padrão do português, as notas em COMP1 mostraram-se decisivas nesse processo de delimitação. Vejamos a figura a seguir:

Figura III-2: Quadro de distribuição (faixa de rendimento em COMP1) da Base-ENEM congelada

Ri-S = faixas de rendimento – Cf. Tab. IV-2

Note-se que embora as faixas “Re” e “S” tenham sido excluídas da base a partir da nota global esse grupo de textos bem avaliados ainda aparecem no quadro acima. Isso se explica, conforme dissemos atrás, pela boa avaliação dos textos no quesito “norma culta” mesmo nos casos de rendimento insuficiente.

Por isso, aplicando-se o princípio da probabilidade de desvios a esse quadro de COMP1, também determinamos a exclusão das faixas “Re” e “S” na variante COMP1, o que resultou num total de 274.497 textos (ou 28,97%) retirados da Base-ENEM congelada.

Guiado por dados objetivos, esse foi o último recorte promovido na fonte do nosso corpus. Com ela configuramos a base de dados que será tomada a partir daqui como a referência imediata do nosso corpus. Denominamos esse referencial de ***Base-CORVO***, cujo percurso de formação pode ser expressado da seguinte maneira:

Tabela III-5: Quadro de formação da Base-CORVO – recortes sob dados objetivos

Especificação	Frequência	%	Nos. gerais
Base-INEP	1.250.834	100,0	1.250.834
RC dados não válidos	119.406	9,6	1.131.428
RC nota global	183.864	14,7	947.564
RC nota COMP 1	274.497	28,97	673.067
Base-CORVO	673.067	46,30	

RC = recorte

O quadro geral da Base-Corvo, com exatas 673.067 unidades e suas associações está expresso na tabela IV.6 a seguir:

Tabela III-6: Base-CORVO – distribuição entre as variáveis do perfil do redator

PERFIL			FAIXA DE RENDIMENTO EM COMPI											
			1 = Ri				2 = Rb				3 = Rm			
	Segmento	Freq.	%	% Acum.		Freq.	%	% Acum.		Freq.	%	% Acum.		
IDADE	1	3.526	10,09	10,09		10.142	14,57	14,57		159.409	28,04	28,04		
	2	19.872	56,84	66,93		42.300	60,78	75,35		324.986	57,16	85,20		
	3	11.562	33,07	100,00		17.151	24,64	100,00		84.119	14,80	100,00		
	Total	34.960	100,00			69.593	100,00			568.514	100,00			
ORIGEM Região (B)	1	AC	93	0,3	Sob (A)	Sob (B)	163	0,2	Sob (A)	Sob (B)	933	0,2	Sob (A)	Sob (B)
					0,3	3,7			0,2	3,0			0,2	2,9
		AM	559	1,6	1,9	22,2	667	1,0	1,2	12,2	6.181	1,1	1,3	19,1
		AP	168	0,5	2,4	6,7	246	0,4	1,6	4,5	1.995	0,4	1,7	6,2
		PA	611	1,7	4,1	24,3	1.934	2,8	4,4	35,4	12.052	2,1	3,8	37,3
		RO	316	0,9	5,0	12,6	770	1,1	5,5	14,1	4.120	0,7	4,5	12,8
		RR	319	0,9	5,9	12,7	682	1,0	6,5	12,5	2.822	0,5	5,0	8,7
		TO	451	1,3	7,2	17,9	1.009	1,4	7,9	18,4	4.199	0,7	5,7	13,0
	Total	2.517	7,2		100,0	5.471	7,9		100,0	32.302	5,7		100,0	
	2	AL	314	0,9	8,1	2,3	674	1,0	8,9	2,7	3.564	0,6	6,3	2,4
		BA	3.133	9,0	17,1	23,2	8.162	11,7	20,6	32,2	42.725	7,5	13,8	28,4
		CE	2.979	8,5	25,6	22,0	6.923	9,9	30,5	27,3	39.074	6,9	20,7	25,9
		MA	2.678	7,7	33,3	19,8	1.168	1,7	32,2	4,6	12.803	2,3	23,0	8,5
		PB	275	0,8	34,1	2,0	1.007	1,4	33,6	4,0	6.675	1,2	24,2	4,4
		PE	1.815	5,2	39,3	13,4	4.715	6,8	40,4	18,6	26.362	4,6	28,8	17,5
		PI	599	1,7	41,0	4,4	1.259	1,8	42,2	5,0	6.105	1,1	29,9	4,1
		RN	1.471	4,2	45,2	10,9	635	0,9	43,1	2,5	8.965	1,6	31,5	6,0
		SE	257	0,7	45,9	1,9	840	1,2	44,3	3,3	4.329	0,8	32,3	2,9
		Total	13.521	38,7		100,0	25.383	36,4		100,0	150.602	26,6		100,0
	3	ES	983	2,8	48,7	6,7	294	0,4	44,7	1,0	12.290	2,2	34,5	4,5
		MG	1.873	5,4	54,1	12,8	6.793	9,8	54,5	23,5	64.783	11,4	45,9	23,5
		RJ	2.097	6,0	60,1	14,4	3.199	4,6	59,1	11,1	35.645	6,3	52,2	12,9
		SP	9.637	27,6	87,7	66,1	18.658	26,8	85,9	64,5	163.366	28,7	80,9	59,2
Total		14.590	41,8		100,0	28.944	41,6		100,0	276.084	48,6		100,0	

PERFIL			FAIXA DE RENDIMENTO EM COMPI											
			1 = Ri				2 = Rb				3 = Rm			
	Segmento		Freq.	%	% Acum.		Freq.	%	% Acum.		Freq.	%	% Acum.	
ORIGEM Região (A)	4	PR	893	2,6	Sob (A)	Sob (B)			Sob (A)	Sob (B)			Sob (A)	Sob (B)
		RS	196	0,6	90,3	47,8	3.646	5,2	91,1	78,5	34.631	6,1	87,0	54,2
		SC	779	2,2	90,9	10,5	928	1,3	92,4	20,0	18.237	3,2	90,2	28,5
		Total	1.868	5,4	93,1	41,7	70	0,1	92,5	1,5	11.082	1,9	92,1	17,3
	5	DF	239	0,7			4.644	6,6			100,0	11,2		
		GO	1.365	3,9	93,8	9,7	613	0,9	93,4	11,9	4.024	0,7	92,8	8,8
		MS	349	1,0	97,7	55,4	1.244	1,8	95,2	24,2	16.262	2,9	95,7	35,7
		MT	511	1,5	98,7	14,2	1.608	2,3	97,5	31,2	15.812	2,8	98,5	34,7
	Total	2.464	7,1	100,0	100,0	1.686	2,4	99,9	32,7	9.478	1,7	100,2	20,8	
		2.464	7,1	100,0	100,0	5.151	7,4	100,0	100,0	45.576	8,1	100,0	100,0	
RENDA	0		1.012	2,89	2,89		1.578	2,27	2,27		7.790	1,37	1,37	
	1		29.238	83,63	86,52		56.227	80,79	83,06		408.952	71,93	73,3	
	2		4.710	13,47	100,0		11.788	16,94	100,0		151.772	26,70	100,0	
	Total		34.960	100,00			69.593	100,00			568.514	100,00		
INSTITUIÇÃO	A		32.012	91,57	91,57		62.085	89,21	89,21		458.212	80,60	80,6	
	B		1.390	3,98	95,55		2.984	4,29	93,5		30.185	5,31	85,91	
	C		1.508	4,31	100,0		4.444	6,39	100,0		79.645	14,01	100,0	
	D													
	Total		34.960	100,00			69.593	100,00			568.514	100,00		

Em Origem, “sob (A)” = porcentagem sobre o total no país “sob (B) = porcentagem sobre o total na região

Com esse panorama cremos que agora o nosso leitor possui o repertório técnico necessário para acompanhar o nosso estudo de corpus.

Neste capítulo estivemos ocupados em “instrumentalizar” o nosso interlocutor não só com as reflexões mais densas da Lingüística de Corpus, como também em mostrar que a delimitação de um corpus é uma tarefa que exige posicionamento conceitual do projetista e tomadas de decisão em favor da confiabilidade do corpus para a análise lingüística. Mas, como veremos no próximo capítulo, essa confiabilidade só é plenamente garantida com uma metodologia explícita de *design* do corpus. Vamos a ela, então.

IV. CORVO – desenho do corpus e modelagem de aplicativos

A metodologia desta pesquisa, como dissemos no início, está apoiada nos estudos com corpus à luz das principais diretrizes da Linguística de Corpus (LC). Além disso, conforme adiantamos na introdução, utilizamos dois instrumentos de apoio metodológico neste trabalho: um sistema de revisão automática da escrita e um aplicativo para a anotação de desvios. O propósito deste capítulo é descrever esse percurso metodológico que adotamos, dando conta da construção propriamente dita do nosso corpus de pesquisa, bem como dessas ferramentas que nos auxiliaram nesse trajeto.

Na seção anterior, nosso leitor foi levado a compreender que esse corpus de estudo se ocupa de um conjunto de produções escritas específico, partindo de uma base textual bastante ampla, que é a Base-ENEM. A fim de que este estudo se desenvolvesse a contento, tornou-se necessário um recorte refinado dessa fonte textual e o resultante é o que denominamos aqui de corpus **CORVO**.

Durante a exposição neste capítulo, abordaremos o Corvo especificamente. Detalhar a sua construção nos pareceu a maneira mais eficaz de apresentarmos o plano metodológico assumido na investigação. Como vimos, há em LC um certo passo a passo já consagrado de construção de corpus, e isso, aqui, foi levado em conta. Ao discorrermos sobre as etapas de construção do Corvo, portanto, nosso leitor poderá, além de refletir sobre as decisões adotadas em cada uma delas, também notar as características que as unidades desse corpus adquiriram gradualmente até ficarem prontas para uma análise linguística relevante.

Quanto aos materiais que serão apresentados neste capítulo, ambos se introduzem numa etapa determinada da construção do Corvo, qual seja, a etapa da anotação. Ver-se-á, desse modo, que a contribuição desses instrumentos está de acordo com o nosso interesse em associar ao trabalho com o corpus um mecanismo automático de modelagem dos textos. Mas isso é uma particularidade desta investigação que, na fase de análise, também procurará dar conta do mérito e da limitação seja do revisor automático seja do aplicativo de edição. Assim, nosso leitor não deve entender que há peso diferenciado entre as etapas de construção de um corpus, mas, antes,

um aprofundamento nosso sobre a questão da anotação por estar diretamente envolvida nos objetivos da investigação.

Esta seção está dividida em duas partes. Na primeira e mais extensa, apresentamos o Corvo e seu processo de construção; na segunda, apontamos o trabalho de preparação e tabulação estatísticas de dados. Esta última vem a ser fundamental para a compreensão da análise que se desenvolverá adiante.

Antes de partirmos para a exposição, cumpre revelar, a respeito do nome Corvo, que se trata de uma denominação abstrata, sem conexão direta com a noção de corpus ou a de desvio. A intenção foi indicar, metaforicamente, a idéia transmitida por E.A. Poe em seu poema “The Raven”, segundo a qual determinadas coisas relegadas a baixo, tomadas como em decadência, escondem, na verdade, propriedades de alto valor. No poema, o corvo incorpora essa atitude de resgate e de ascensão do imprestável, do inoportuno, do desprezível. Este é, justamente, o movimento que pretendíamos com o nosso trabalho de corpus, fundado em textos com provável alta frequência de desvios em relação ao padrão da escrita culta.

IV.1 CORVO – o corpus

O processo de construção do Corvo obedeceu às etapas a seguir. Mas ressaltamos, por pertinente, que esse processo é um *movimento cíclico* para o pesquisador, construtor do corpus; de modo que os reajustes, naturais e freqüentes, levam-no a saltar etapas ou retornar às iniciais, às vezes com um pequeno grupo, outras com o conjunto inteiro dos textos. O que temos, enfim, é um “roteiro de trabalho” consistente em seus critérios conceituais com fases específicas de tratamento dos textos. São estas etapas:

- a) compilação: estudo da fonte do corpus, seleção das amostras (amostragem) e apresentação dos textos em formato eletrônico, válido para as tarefas às quais se destinam;
- b) armazenamento: organização da estrutura do corpus, alocação das amostras e padrão da nomeação de arquivos compatível com a estrutura geral do corpus e com mecanismos eficientes de busca automática;
- c) documentação: tratamento dos dados externos dos textos (tais como os de autoria), por meio de estratégia automática de associação de dados, permitindo, dentre outros, a geração de um banco

de dados para mapeamento e recuperação das informações relativas ao perfil das amostras do corpus;

d) anotação: inserção de informações lingüísticas, de caráter técnico, com vistas a informar, pela via automática, determinadas propriedades lingüísticas dos textos ou de segmentos específicos deles.

Há, além dessas, uma importante etapa do processo de construção de um corpus, que é a disponibilização. Não cumprimos essa fase com o Corvo, porque esse é um corpus de estudo e temos restrições legítimas de disponibilização pública. Mas reconhecemos a importância do trabalho que se realiza com a disponibilização de um corpus, e talvez haja fôlego, fora do escopo deste trabalho, para perseguirmos essa meta com o Corvo. Retornaremos a essa questão nas nossas considerações finais. A partir daqui, discorreremos sobre as quatro fases efetivamente realizadas no desenvolvimento do corpus.

IV.1.1 COMPILAÇÃO

Como se sabe, até aqui o material de base para a formação do nosso corpus – redações do ENEM-2002 – sofreu análises e manipulação de dados para que pudéssemos nos concentrar num extrato específico de textos e perfil de redator. Formamos, então, uma base de dados (textos com perfil de redator associado) de partida para o nosso estudo. Essa base recebeu o nome de **Base-Corvo** e reuniu **673.067** textos (cf. Tab. III.6 no capítulo anterior).

O corpus Corvo enfoca essa Base e passa a ser compilado seguindo duas ações: a *seleção textual*, i.e., o recorte que resultará no conjunto textual efetivamente tratado, e a *transposição de registro*, ou seja, a formatação dos textos para serem tomados como unidades de um corpus eletrônico de fato. Na seqüência, discorreremos sobre cada uma dessas ações separadamente.

IV.1.1.1 Seleção textual

Antes de avançarmos na exposição, convém resgatarmos um ponto importante sobre a seleção textual a fim de que nosso leitor tenha clareza do percurso assumido até aqui.

Pelo tratamento dado à base original de dados, seguindo os interesses particulares de nossa pesquisa, sabemos que a população que o Corvo representa é, de partida, apenas uma fatia

do universo do qual participa. Ou seja, falamos de um universo que é o ENEM, edição de 2002, do qual extraímos uma faixa de redatores/textos, qual seja, a de pior desempenho no quesito norma culta (essa faixa compreende, justamente, a Base-Corvo). Para efeito de representatividade, portanto, devemos ter a exata dimensão do recorte que esta investigação promove sobre o material de base.

Outro aspecto relevante sobre esse tema é o fato de estarmos diante de um banco de dados que caracteriza essa população quanto à idade, origem, renda e tipo de instituição na formação escolar. Ora, este trabalho não se ocupa da análise sociológica ou sócio-lingüística do ENEM. Mas, sobre o extrato que recorta desse Exame, temos o interesse de refletir no corpus o comportamento de perfil ali observado. Isso implica dizer, então, que para consolidar a composição do Corvo não nos preocupamos com a sua representatividade em cada um dos fatores que caracterizam os redatores, mas, antes, com uma representatividade suficiente da população da Base-Corvo. Para efeito de análise, como veremos mais tarde, esses dados do perfil serão colocados como *indicativos* das características dos redatores com os quais nos ocupamos. E servirão para nos mostrar caminhos, em vez de funcionarem como elemento motor da investigação.

Dito isso, vamos prosseguir anunciando, de imediato, que estipulamos para o nosso corpus um número de **300** textos. Inicialmente esse número fora menor (em torno de 100) em virtude do escopo desta pesquisa. Contudo, julgamos fundamental o trato de um conjunto maior de amostras para conseguir uma representatividade que tornasse possível a análise do perfil. O esforço que suportamos, embora se desejasse maior, foi mesmo o de aumentarmos em duas vezes os planos de trabalho iniciais e, desse modo, garantir um quadro que, no tocante ao perfil de redatores, fosse significativo.

Fixado o número de 300 unidades a serem extraídas da Base-Corvo, partimos para o levantamento de uma *amostra de qualidade* seguindo procedimentos estatísticos comuns. Detalhemos brevemente este aspecto do trabalho, sobretudo para conquistar do nosso leitor a sua confiança no corpus que será composto a partir dessa amostragem.

Uma amostra de qualidade é aquela que possui as mesmas características da população e que pode representá-la bem, ou seja, é uma amostra que produz resultados muito semelhantes aos da população. As principais dificuldades na aplicação das etapas do levantamento amostral estão

no momento de decidir o tamanho da amostra e como selecionar a amostra da população. Uma boa amostra será obtida dependendo basicamente do número de observações amostradas e do método com que essas observações foram selecionadas. Uma amostra é *estatisticamente representativa*, quando, para uma certa característica, as distribuições da amostra e de sua respectiva população forem iguais, caso contrário diz-se que a amostra é não-representativa, e estudos baseados nessa amostra podem ser enganosos. Uma maneira muito segura de se obter uma amostra representativa é através do processo aleatório de seleção dos elementos.

Numa amostra aleatória, utiliza-se um sorteio aleatório dos elementos da população para fazer sua seleção, podendo ser obtida por diversos métodos, sendo um dos mais comuns a amostragem aleatória simples. Vale dizer que esse e outros métodos (a amostragem estratificada, a amostragem por conglomerados e a amostragem sistemática) estão incluídos numa amostragem probabilística. Vejamos qual o princípio que as reúne aqui e que a torna pertinente para o nosso caso.

Na *amostragem probabilística* reúnem-se todas as técnicas que usam mecanismos aleatórios na seleção dos elementos da amostra, atribuindo a cada um deles uma probabilidade, conhecida *a priori*, de pertencer à amostra. Para tirar conclusões válidas sobre a população de estudo a partir dos resultados da amostragem, a maneira estatisticamente correta de escolher os indivíduos da população seria mesmo por amostragem probabilística. Neste caso, vejamos o que implica adotarmos o método de amostragem de que falamos acima.

Na *amostragem aleatória simples* cada indivíduo da população tem a mesma chance de ser incluído na amostra. Pode-se também dizer que a amostragem aleatória simples é um processo de obtenção de várias amostras tiradas da população em estudo, de modo que cada uma das amostras tenha a mesma chance de ser selecionada. Para uma população de tamanho N e uma amostra de tamanho n , existem C_n^N possíveis amostras, sendo que cada amostra tem uma chance de $1/C_n^N$ de ser selecionada. Ela pode ser realizada com ou sem reposição, mas, na prática, usa-se a amostragem sem reposição. A amostragem aleatória simples pode ser realizada, atribuindo a cada membro da população um único número obtidos por tabelas apropriadas ou mecanismos computacionais.

Encerremos essa digressão conceitual aqui. Em síntese, o que fizemos foi partir da nossa população – 673.067 – para atribuir a cada unidade um número aleatório; em seguida, ordenamos

esse conjunto (que é a Base-Corvo) por esse número e selecionamos os primeiros 300 casos como a nossa amostragem.

A fim de que se verifique a proporcionalidade entre a amostragem (daqui por diante A300) e o material de origem, trazemos abaixo duas tabelas de cada um dos casos, selecionadas pela pertinência (caso da tabela referente à nota em COMP1) e pela abrangência (caso da tabela com os dados sobre a origem do redator)⁴¹.

Tabela IV-1: Distribuição em nota COMP1 da Base-Corvo

Nota COMP1	Frequência	%	Freq. Acum.	% Acum.
1	34.960	5,19	34.960	5,19
2	69.593	10,34	104.553	15,53
3	568.514	84,47	673.067	100,00

1 = Ri 2 = Rb 3 = Rm

Tabela IV-2: Distribuição em origem-região da Base-Corvo

Origem-região	Frequência	%	Freq. Acum.	% Acum.
1	40.290	5,99	40.290	5,99
2	189.506	28,16	229.796	34,14
3	319.618	47,49	549.414	81,63
4	70.462	10,47	619.876	92,10
5	53.191	7,90	673.067	100,00

1 = Norte 2 = Nordeste 3 = Sudeste 4 = Sul 5 = Centro-Oeste

Tabela IV-3: Distribuição em nota COMP1 de A300

Nota COMP1	Frequência	%	Freq. Acum.	% Acum.
1	13	4,33	13	4,33
2	29	9,67	42	14,00
3	258	86,00	300	100,00

1 = Ri 2 = Rb 3 = Rm

Tabela IV-4: Distribuição em origem-região de A300

Origem-região	Frequência	%	Freq. Acum.	% Acum.
1	22	7,33	22	7,33
2	85	28,33	107	35,67
3	133	44,33	240	80,00
4	34	11,33	274	91,33
5	26	8,67	300	100,00

1 = Norte 2 = Nordeste 3 = Sudeste 4 = Sul 5 = Centro-Oeste

⁴¹ Seria possível a mesma conclusão se outras fossem as tabelas (p.ex., referentes à idade, à renda). Mas a intenção aqui é tão somente a de demonstrar a paridade entre a amostragem e a fonte dos dados.

Uma surpresa do processo de investigação foi o fato de, ao definirmos esse material e estarmos prontos para fase seguinte, que é a transposição de registros, termos perdido amostras. Tal acontecimento é comum num trabalho com corpus, sendo previsto nesta etapa de construção em projetos de grandes proporções, a fim de que, com suficiente retaguarda, essas amostras sejam repostas pelos desenvolvedores do corpus. Num corpus de estudo, no entanto, o que observamos é que a falta de reposição é uma espécie de “contingência com efeitos”, i.e., a investigação é passível de imprevistos, mas será preciso levar isso em conta na análise sempre que o grau de interferência do imprevisto se impuser no estudo em questão.

A perda de amostras se deu pela *corrupção de arquivos*, como é conhecido em computação. Um arquivo corrompido é aquele que: a) não abre na plataforma em que foi criado (p.e., um arquivo .doc não abre na plataforma Word); b) não apresenta a integralidade dos dados (p.e., numa planilha em Excel, parte dos dados estão ausentes, embora originalmente editados); ou c) parte dos dados são transformados em códigos, por incapacidade operacional de leitura correta do formato em que foram editados (p.e., num processo de escaneamento de texto verbal, símbolos como aspas ou acentos são transformados em códigos não-verbais por falha na leitura do original impresso). No nosso caso, os arquivos requeridos sequer foram encontrados em algumas vezes; mas de modo geral, os arquivos são corrompidos porque não abriram de jeito algum.

Tivemos uma baixa de 51 amostras, e passamos, então, à compilação textual de **249 textos** efetivamente. O Corvo começa a ser formatado; passará da versão manuscrita à versão eletrônica, como veremos a seguir. Mas, antes, convém apresentarmos o quadro de perfil relativo a esse grupo resultante da amostragem até aqui apresentada.

As 249 amostras foram mapeadas em função de cada um dos fatores que as caracterizam. Apontamos, a seguir, um conjunto de três *tabelas de distribuição* relacionadas a cada um desses fatores. Note-se que essa distribuição já está relacionada aos segmentos do banco de dados separados em função do desempenho em COMP1, i.e., relacionada às três faixas de rendimento aqui observadas: o rendimento insuficiente (Ri), o rendimento baixo (Rb) e o rendimento médio (Rm).

Distribuição pelo fator Idade

Tabela IV-5: Distribuição de A300 por idade

A300	Idade			
	Até 18 anos	18 a 21 anos	Acima 21 anos	Total
Ri	5	1	5	11
Rb	3	19	6	28
Rm	62	117	31	210
Total	70	137	42	249

Figura IV-1: Quadro de distribuição de A300 (percentual por idade)

Figura IV-2: Quadro de distribuição de idade de A300 (percentual por segmentos de A300)

Distribuição pelo fator Origem-região

Tabela IV-6: Distribuição de A300 por origem-região

A300	Idade					Total
	Norte	Nordeste	Sudeste	Sul	C.Oeste	
Ri	0	3	8	0	0	11
Rb	4	8	12	3	1	28
Rm	8	59	100	27	16	210
Total	12	70	120	30	17	249

Figura IV-3: Quadro de distribuição de origem-região de A300 (percentual por origem-região)

A300	Idade					Total
	Norte	Nordeste	Sudeste	Sul	C.Oeste	
Ri	0	4	7	0	0	4
Rb	33	11	10	10	6	11
Rm	67	84	83	90	9	84
Total	100	100	100	100	100	100

Figura IV-4: Quadro de distribuição de origem-região de A300 (percentual por segmentos de A300)

A300	Idade					Total
	Norte	Nordeste	Sudeste	Sul	C.Oeste	
Ri	0	27	73	0	0	100
Rb	14	29	43	11	4	100
Rm	4	28	48	13	8	100
Total	5	28	48	12	7	100

Distribuição pelo fator Renda

Tabela IV-7: Distribuição de A300 por renda

A300	Renda			Total
	Nenhuma	Até 5sm*	Acima de 5sm	
Ri	0	9	2	11
Rb	1	24	3	28
Rm	7	158	45	210
Total	8	191	50	249

*sm = salário mínimo

Figura IV-5: Quadro de distribuição em renda de A300 (percentual por renda)

A300	Renda			Total
	Nenhuma	Até 5sm	Acima 5sm	
Ri	0	5	4	4
Rb	13	13	6	11
Rm	88	83	90	84
Total	100	100	100	100

Figura IV-6: Quadro de distribuição em renda de A300 (percentual por segmentos de A300)

Distribuição pelo fator Instituição

Tabela IV-8: Distribuição de A300 por instituição

A300	Instituição				Total
	Somente Púb.	Maior parte Púb.	Somente Part.	Maior parte Part.	
Ri	11	0	0	0	11
Rb	26	0	2	0	28
Rm	174	10	26	0	210
Total	211	10	28	0	249

Figura IV-7: Quadro de distribuição em instituição de A300 (percentual por instituição)

Figura IV-8: Quadro de distribuição em instituição de A300 (percentual por segmentos de A300)

Esse grupo de amostras também é caracterizado pelas notas recebidas no ENEM nas cinco competências avaliadas; há também a nota global de cada texto. Abaixo, apresentamos essa caracterização em termos de *média* das notas em cada um dos segmentos observados:

Figura IV-9: Distribuição das notas em cada competência nas faixas de rendimento de A300

Esse “mapeamento” dos dados estará em foco na análise de corpus propriamente dita. Agora, sim, passemos à compilação textual.

IV.1.1.2 Transposição de registros

Nos tempos atuais, é comum que o corpus seja compilado diretamente em meio eletrônico: os textos-fonte já estão “digitados”. Para nós o desafio maior desta etapa foi, justamente, a *transposição de registro*, i.e., passar os textos do formato manual para o formato eletrônico. Apresentamos, a seguir, dois aspectos importantes desse processo: o primeiro, em que discorreremos sobre as decisões de projeto envolvidas nessa atividade de transposição, e o segundo, em que abordamos o trabalho em termos da anotação. Este último está deslocado do tópico de discussão específica (deveria constar da etapa de anotação), mas preferimos tratá-lo aqui a fim de que nosso leitor tenha uma visão linear da compilação textual que desenvolvemos.

IV.1.1.2.1 Digitação vs. Transposição de registro

No nosso trabalho, não falamos estritamente de digitação, porque encontramos o texto-fonte (TF) digitalizado. Ou seja, o que há é uma *imagem* da redação do ENEM, conforme exemplificamos abaixo⁴²:

⁴² Na ilustração, vê-se apenas um segmento da imagem.

Figura IV-10: Ilustração do Texto-Fonte

0535179-0 2 B (N) D 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

23010.65.002.047
**FOLHA DE REDAÇÃO
ENEM 2002**

23010

0535179-0 3 B (N) D 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

A cidadania não é exercida por todos, pois nem
aquele que esta palavra significa e descrevem seus próprios
deveres como eleitor, pois alguns não importam por uma
classe minoritária e dominante que utilizam discursos anti-
semanter no poder e iludem aqueles desprivilegiados e misse

Como se vê, TF não é um texto impresso; é, antes, um texto escrito manualmente, do qual temos uma “foto”. Com isso, a qualidade da leitura não é a mesma de um processo de digitação comum. Soma-se, ainda, o fato de que essas fotos não focalizaram apenas a caixa de texto, o que reduziria os problemas de leitura; ao contrário, quase sempre foi incluído o cabeçalho por inteiro nessa fotografia. Assim, podemos dizer que estamos lidando com um *registro* lingüístico – a imagem do manuscrito – e buscamos transformá-lo em outro: um texto eletrônico. E há, nesse processo, um ingrediente adicional, que é a interpretação do especialista (digitador) sobre o que está escrito.

Os milhares de textos do ENEM variam entre si na grafia e no *layout*, o que é próprio do registro manuscrito. Para nós, ao lidarmos diretamente com essa questão, tomamos decisões buscando uma “padronização” desse modo de *mostrar* a linguagem, i.e., a apresentação dos textos eletrônicos o mais possível fiel ao original, ainda que ocorrências muito particulares de um ou outro texto não fossem transpostas, contando, para isso, com critérios de observação e análise de rejeição previamente estabelecidos.

Veja-se um fato que ilustra esse tipo de decisão. Compreender que na transposição de registro não se interpreta somente a grafia, mas uma diversidade de informações possíveis (e mesmo comuns) dentro do manuscrito não nos afasta do problema de que não há absoluta correspondência entre os recursos do manuscrito e os do registro eletrônico escrito. Uma flecha, por exemplo, que no manuscrito indica segmentos textuais que devem ser lidos na direção

apontada, é um tipo de informação que não se reproduz sem desafios num editor de texto. Num caso desses, o leitor tem de tomar decisões e, como resultado teremos um texto (eletrônico) que será, sempre, uma *interpretação* (do manuscrito). Ora, de tudo que pode ser “lido” do TF, o digitador do Corvo esteve preparado para apreender apenas uma parte. Tomemos o exemplo da flecha, novamente. A informação lida com esse indicador não foi reproduzida no texto do corpus. Isso porque buscamos um texto corrido, linear, efetivamente “escrito”. Temos, aqui, um desses critérios de observação com decorrente rejeição de transposição.

Na *formatação* do texto cru, dois padrões foram empregados:

- ✍ operacional: texto em .txt; uma linha entre os parágrafos; sem tabulação; uma linha entre título e corpo do texto.
- ✍ lingüístico: parágrafos com iniciais maiúsculas; todas as marcas de início de parágrafo mantidas (ex.: travessão, ponto, etc); apenas recursos do código verbal mantidos.

Ambos os padrões são verificados em trabalho com corpus. São tomados em virtude de sua eficácia computacional – por exemplo, é mais confortável manipular textos em .txt, ainda que recursos de formatação desejáveis sejam perdidos (negrito e itálico, p.ex.). No aspecto lingüístico-computacional, isso se justifica em vista de que, à custa de uma intervenção indesejável no texto, inserir uma linha entre o título (quando houver) e o corpo do texto, pode, no futuro, beneficiar a disponibilização e o acesso ao Corvo, tornando possível, por exemplo, extrair subcorpora focados no título dos textos. A inserção do título no *cabeçalho* é outro exemplo de benefício. Uma formatação textual atenta ao título é um facilitador no momento de modelagem de um mecanismo de inserção automática de informações no cabeçalho dos textos.

De todo modo, é preciso notar que a intervenção do especialista-digitador nesse tipo de transposição de registros pondera duas tendências. A primeira é a de informar alguma coisa que o texto original não tem, por exemplo, a quebra de linha do título quando não é originalmente o caso. A segunda é a de não informar aquilo que o texto original tem, por exemplo, que determinado segmento estava sublinhado. Trata-se, porém, de agir com parcimônia na relação com essas limitações do campo operacional bem como do lingüístico. O custo-benefício do texto sem formatação e a vantagem na disponibilização do corpus com a segmentação textual (título) foram decisões do Corvo, e o seu usuário deverá sabê-lo.

A par dessas considerações e critérios, podemos dizer que os textos selecionados pela amostragem (249 unidades) foram transpostos efetivamente e constituem, agora, um *corpus*. Suas amostras não são o texto original, também não são TF, mas são, isso sim, o primeiro texto do corpus, a versão base, o texto do corpus mais próximo do original.

Nessa versão, o texto é completo, i.e., foi reproduzido na íntegra. Para efeito de classificação em gênero, são “produções escolares”; na verdade, um tipo específico de produção escrita escolar – uma *redação*: tarefa programada, realizada sob condições singulares e que representa, no mais, “um fato” apenas na vida do redator do português brasileiro escrito. No entanto, esses textos são de extrema importância para o Corvo, posto que constituem textos “não revisados”, tomados, assim, como produções genuínas de certo tipo de redator.

Houve, no entanto, um fato novo que mereceu atenção, aprofundamento e que transformou essa atividade de transposição em algo mais complexo nesta pesquisa. Trata-se de um aspecto da “anotação de corpus”, de que falamos a seguir.

IV.1.1.2.2 Anotação da superfície textual

Como dissemos atrás, o nosso TF é uma foto do texto manuscrito. Mas para a transposição de registro, nos preocupou o fato de que o usuário do Corvo (assim como nós) não tem acesso nem ao texto original nem ao TF, e talvez quisesse observar determinados aspectos do manuscrito ou, ao menos, ser informado sobre alguns traços peculiares, mas recorrentes no corpus. Concentramo-nos, assim, naquilo que denominamos *superfície textual*.

A superfície do texto é, para nós, aquilo que envolve ou está na fronteira fluida entre o simples *layout* do texto e o processo propriamente dito de produção da escrita. Supondo, então, que a “fidelidade” do corpus passa pela questão de reproduzirmos algumas facetas da superfície textual e, uma vez que o digitador é aquele que lida diretamente com TF, decidindo o que transpõe ou não para a versão eletrônica do texto, resolvemos investir na *anotação (ou etiquetagem)* para tratarmos essa questão.

Ainda neste capítulo, abordaremos detidamente a anotação de corpus promovida no Corvo. Mas, neste momento, cumpre dizermos que a anotação foi o mecanismo assumido para indicarmos determinadas propriedades da superfície dos nossos textos. Isso porque aproveitamos

uma tecnologia eficaz – que é a XML – para isolarmos, se quisermos, as ocorrências desse tipo; ou seja, como se trata de uma intervenção não textual, era desejável que elas pudessem ser ocultadas, desconsideradas ou mesmo eliminadas, e isso pela via automática. A inserção desse tipo de marca no texto, através das etiquetas em XML, também nos mostrou satisfatória, porque poderíamos distinguir os fatos da superfície textual observados, criando condições de mapeamento mais refinado dessa superfície no corpus como um todo. Por fim, a possibilidade de contabilizarmos esse trabalho oferecida pela anotação em XML foi determinante.

Com clareza sobre a opção por esse suporte, votemos à questão da superfície textual, recuperando algo que acabamos de discutir, que é o fato de um texto manuscrito ser repleto de informações muitas vezes não reproduzidas na transposição para a versão eletrônica. Se no tratamento do extra-texto verbal houve perda de informações transpostas (pensemos no caso da flecha, do negrito e itálico, p.ex.), desta vez nossa preocupação foi a de enriquecer o texto eletrônico com informações tipicamente não transpostas. Ocorreu, em outras palavras, o máximo de “aproveitamento” daquilo que, em geral, são tomados como acidentes da escrita não-revisada, tais como o apagamento do final de uma palavra. É bem verdade que aspectos ligados à caligrafia, por exemplo, não foram observados (letra inclinada vs. letra arredondada; alternância no uso de letras maiúsculas e minúsculas; alternância entre letra cursiva e letra bastão; etc.). Mas foi possível tratar outros fatos de relevância lingüística ou lingüístico-computacional.

Esses fatos, no Corvo, estão separados nas categorias apontadas a seguir. Cada um acompanha-se da etiqueta correspondente e suas variações. Apontamos, aqui, os casos em que as etiquetas foram aplicadas, seguindo-se a sintaxe padrão do XML, junto à ocorrência detectada:

a) qualificação do título – `<tit_rep>`; `<tit_pt>` : quando o texto apresenta como título a reprodução do tema indicado para a redação. Mas quando há título e esse não repete o tema da redação, ele não recebe etiqueta. Por sua vez, os títulos, em geral, não têm pontuação final. Mas para os casos em que o ponto final (pt) foi empregado, e somente nesses casos, anotou-se a ocorrência com a etiqueta específica.

b) ilegibilidade pontual – `<ileg>`; `<ileg_lex>`; `<ileg_gr>`: quando determinado segmento textual não oferece condições de decifração. Criamos três tipos de etiquetas de ilegibilidade: a primeira que é geral; a segunda para os casos de ilegibilidade de uma palavra; e a terceira para as ocorrências de ilegibilidade ligadas à estrutura da frase (conectivos e pontuação, por exemplo).

c) apagamento – <ap>; <ap_lex>; <ap_sbr_lex>: quando determinada sequência (de letras ou palavras) foi propositalmente ocultada pelo redator e representa um apagamento linguisticamente relevante. Usamos a etiqueta geral para o apagamento de letras e de sequências não tomadas como palavras, e a etiqueta específica para os casos de apagamento lexical propriamente dito. A etiqueta com “sbr” foi criada no percurso da anotação. Indica, para nós, os casos de “apagamento com sombra”; ou seja, quando houve uma tentativa de ocultar o segmento, mas ela não teve sucesso, o que nos permite saber (mas não reproduzir) o trecho ocultado.

d) inserção – <in_letra>; <in_cadeia>; <in_lexgr>: quando o redator introduz determinado segmento no texto já produzido e essa inserção altera substancialmente o segmento textual a que pertence. Para esta anotação, distinguimos a inserção de letra, a inserção de uma sequência de palavras e a inserção de um item léxico-gramatical (uma conjunção, por exemplo).

e) hesitação – <h_ort>; <h_ac>; <h_lexgr>; <h_num>: quando um desvio gramatical seria cometido, mas o redator o detectou a tempo, corrigiu-o, deixando uma marca explícita de sua hesitação. A anotação que criamos separa os casos de hesitação na ortografia, na acentuação, no uso de elementos léxico-gramaticais (expressões fixas, crase, p.ex.) e no emprego da concordância em número.

f) cabeçalho – <cab>: quando são indicados local e data, por exemplo, no topo do texto.

g) assinatura – <ass>: quando o texto é assinado.

h) ruído – <r>: quando determinado segmento do texto tem a leitura prejudicada por algum tipo de contingência, que não pode ser considerada um apagamento ou uma hesitação; é o caso de uma borra de caneta ou mancha do papel, p.ex..

Temos, com isso, oito categorias de etiquetas de superfície, com variação em algumas delas. Todos os textos do Corvo foram observados quanto a esses aspectos e, quando foi o caso, receberam as etiquetas pertinentes. A seguir, ilustramos essa anotação reproduzindo o trecho das redações em que ela aconteceu:

Tabela IV -9 : Ilustração da anotação da superfície textual

Superfície Textual	Critério	Trecho
Título	Reprodução do tema	<tit_rep>O direito de votar: como fazer dessa conquista um meio para promover as transformações sociais de que o Brasil necessita?</tit_rep> O direito de votar é muito importante, pois toda pessoa que tem idade (...) [C3C_43]
	Término com ponto final	O direito de votar<tit_pt>.</tit_pt> O meu voto, é muito iportante para escolher (...) [C2A_1]
Ilegibilidade	Nenhuma decifração	(...) eu acho que ele seja a conquista de um povo que não tinha o direito nem de escolher quem <ileg></ileg> ser o seu representante mundial. [C3C_92]
	Decifração insatisfatória	O voto é para se saber qual é o verdadeiro <ileg_lex>agente</ileg_lex> que colocamos para governa nossas (...) [C2C_12]
Apagamento	Total em palavra	(...) não da boca prá fora mas sim com trabalho digno e honesto, <ap_lex></ap_lex> merecedor do seu voto, pois se ver muita falsidade (...) [C3C_77]
	Parcial em seqüência de palavras	Dessa forma as pessoas não <ap_sbr_cadeia>pensam e</ap_sbr_cadeia> análise os candidatos (...) [C3C_36] Vamos a luta e mostrar <ap_sbr_lex>que</ap_sbr_lex> <ap_lex></ap_lex> respeito. [C3C_54]
Inserção	Inserção de letra	Em tempos de políticas, é o tempo em que nós nos manifestamos contras os <in_letra>u</in_letra> maus politicos (...) [C2C_24]
	Inserção de palavra	Mas uma obrigação que agente <in_lex>tem</in_lex> com o nosso País. [C3C_96]
	Inserção de componente estrutural	A partir do momento <in_lexgr>em</in_lexgr> que o povo Brasileiro adquirio esse direito, de decidir (...) [C2C_20]
Hesitação	Reformulou a ortografia	O mundo em que vivemos nos <h_ort>tráz</h_ort> trás muitos riscos e dificuldades (...) [C4C_12]
	Repensou o emprego do hífen	(...) pois se o país tem a chance de mudar, basta <h_hf>aproveitála</h_hf> aproveitá-la muito bem. [C4C_25]
	Repensou o emprego do acento	A escolha é muito importante, pois é <h_ac>está</h_ac> esta que ditará as regras da eleição, quando (...) [C3C_29]
Cabeçalho	Apresenta cabeçalho	<cab>Salvador-BA 25 de Agosto de 2002.</cab> Para a direção deste concurso a minha opinião sob o tema "E". O Brasil necessita (...) [C2A_2]
Assinatura	Apresenta assinatura	O Brasil já necessita de fazer grandes transformações em vários pontos que nosso povo vive. isto é tudo <ass></ass> [C3C_78]
Ruído	Presença de falhas	(...) e não levar para o <r>país</r> fundo do poço o país, inteiro, valorize voto. [C2C_27]

Alguns casos desse exemplário nos permite abordar certas especificidades da anotação. Observe-se, por exemplo, que algumas vezes o componente textual não aparece na sintaxe da etiquetagem. Empregamos o traço subscrito () para esses casos e, se notarmos, isso acontece nos casos de ilegibilidade ou apagamento, mostrando que não foi possível ler o que estava escrito, ou nas ocorrências de assinatura, em que é impraticável a sua transposição.

Pode-se perceber, de modo geral, que tanto a posição da etiqueta como o preenchimento da sintaxe podem variar conforme o fato de superfície observado. Por exemplo, nas etiquetas de hesitação e na de inserção de letra o que se tem é a transposição de um segmento textual que, na verdade, deveria ter sido descartado. Por isso, ao manipular esses casos, será preciso saber que essas etiquetas especialmente têm ligação com o segmento textual imediato. Por exemplo, na ocorrência de hesitação em C4C_12, a reformulação em questão recai sobre “trás”; na ocorrência de inserção de letra em C2C_24, a letra “u” foi inserida na palavra “maus”; e assim por diante.

Importante comentar, por fim, a variação que a etiquetagem da hesitação apresenta. É, talvez, o tipo de etiqueta com maior diversidade de especificação, e isso se justifica pela incapacidade de prevermos com qual tipo de hesitação podemos nos deparar no processo de leitura e transposição textual. Por isso, vêm-se, acima, hesitação recaindo sobre a ortografia, sobre a acentuação, sobre a aplicação de hífen, enfim, um elenco de fatos lingüísticos que se distinguem entre si. Por se tratar de algo bastante peculiar desta pesquisa, optamos por não controlar essa lista de especificações, tornando a etiquetagem da hesitação algo para ser refinado a partir da própria análise do corpus.

Conduzida dessa maneira a transposição de registro, pode-se dizer, a essa altura, que a primeira versão das amostras do Corvo são textos já anotados. Não existe, portanto, o chamado “texto cru”, como é típico de um corpus em grandes projetos, em que se tem uma versão base dos textos sem qualquer marcação. Haverá, certamente, casos de nenhuma ocorrência dos fatos de superfície textual aqui observados. Porém, isso já está no âmbito da análise e não da construção do corpus, que é sobre o que nos concentramos neste momento.

IV.1.1.3 ARMAZENAMENTO

Uma vez compilados, é preciso guardar os textos adequadamente no computador. Nesta etapa, são os arquivos, e não propriamente os textos, o foco da atenção. Por isso, as principais decisões desta fase são ligadas à nomeação de arquivos e à estruturação geral do corpus.

IV.1.1.3.1 Nomeação de arquivos

Para nomear os arquivos do Corvo consideramos, por um lado, o imperativo em LC para se criar nomes curtos em favor da posterior manipulação dos arquivos por ferramentas de processamento de corpus. Por outro lado, nossos textos têm uma autoria minimamente caracterizada, e desejamos que esse perfil se refletisse no nome dos arquivos, até mesmo para, num futuro, facilitar a modelagem dos mecanismos de busca e acesso ao corpus.

Para satisfazer essas premissas, analisamos os dados do perfil, concluindo que o traço da origem (região) nos oferecia as vantagens de: a) ser exaustivo, i.e., é uma informação preenchida em todos os casos, não só das amostras do corpus, mas do ENEM como um todo; b) nos mostrar um mapeamento geográfico do corpus, apresentando um quadro nacional mais abrangente do que se optássemos pelo mapeamento pela idade, por exemplo; e c) haver maior variação dentre os itens do perfil; ou seja, enquanto a origem é um dado subespecificado em 5 classes (cada uma das regiões do país), os dados relativos à renda, por exemplo, são subespecificados em apenas 3.

Com isso, modelamos a nomeação dos arquivos, conforme a ilustração abaixo:

Figura IV-11: Modelo de nomeação de arquivos do Corvo

Nome	Expressão
<p><i>Matriz</i></p>	<p>Base de referência: cód.: C = Corvo</p> <p>Região (perfil do redator): código numérico, significando: 1 = região Norte, 2 = Nordeste, 3 = Sudeste, 4 = Sul, e 5 = Centro-Oeste.</p> <p>Faixa de rendimento (perfil do desempenho no ENEM): código alfabético, significando: A = Ri, B = Rb, e C = Rm</p> <p>Número de ordem dentro do diretório a que pertence</p>
<p><i>Exemplos</i></p> <p>C2A_1</p> <p>C3C_94</p>	<p>Corvo Ri (rend. insuficiente)</p> <p>Região Nordeste 1º arquivo desse perfil</p> <p>Corvo Rm (rend. médio)</p> <p>Região Sudeste 94º arquivo desse perfil</p>

Note-se que:

a) o que existe é um *código* efetivamente, que alterna letras e números e isso numa variação controlada (apenas o número de ordem é que se tem em aberto);

b) cuidamos para que o modelo fosse flexível a ponto de permitir, num futuro, nomear outros corpus a partir da Base-ENEM, distinguindo as amostras, por exemplo, de acordo com outros itens do perfil.

IV.1.1.3.2 Estrutura do corpus

A organização geral do corpus se dá em diretórios e subdiretórios (ou pastas). Para nós, o topo dessa hierarquia é a Base-Corvo. Os membros, por sua vez, dividem-se nas faixas de rendimento (R_i , R_b e R_m), e esses em cada uma das regiões geográficas.

Assim, em termos gerais, o Corvo apresenta: um diretório, dividido em 3 subdiretórios e estes em 5 pastas. Os arquivos são o mais baixo nível da estrutura. Veja-se tal configuração na ilustração abaixo:

Figura IV-12: Estrutura geral do Corvo

Essa estrutura do corpus oferece vantagens para a tabulação de dados. Mas, acima disso, tal estrutura pode ser bastante útil sob dois aspectos. O primeiro é que ela pode contribuir na elaboração do programa de busca do Corvo, já que separa os textos em categorias (de perfil e de rendimento) que são muito claras. O segundo é que essa estrutura permite expandir o corpus sem grandes ajustes no plano inicial de organização.

IV.1.1.4 DOCUMENTAÇÃO

As amostras do corpus, devidamente formatadas e armazenadas, passam pelo processo de *documentação*, que consiste em indexar ao texto informações externas. Isto é, criamos o chamado *cabeçalho*.

Para o Corvo, projetamos um cabeçalho que é rico, diferencial e operacionalmente simples. É rico, porque contém dados sofisticados sobre a autoria, que são justamente as informações trazidas do ENEM e tabuladas em nosso banco de dados. É diferencial, porque relaciona o texto às notas recebidas no ENEM. E, finalmente, é operacionalmente simples, porque as informações podem ser resgatadas automaticamente (com o intercâmbio com o banco de dados) e porque, apesar do volume, boa parte do corpus tem dados identificadores comuns a todas as amostras (p.ex., a fonte, que é o ENEM se repete em todos os casos).

Nosso modelo de cabeçalho apresenta os dados separados em 5 itens de identificação no corpus, conforme se vê no esquema abaixo:

Figura IV-13: Modelo de cabeçalho do Corvo

FICHA 1 = <i>CORPUS</i> ✍Endereço (do arquivo no corpus) ✍Categoria (do corpus) ✍Língua ✍Tokens (no.) *Categoria: Desvio *Língua: Português Brasileiro	FICHA 2 = <i>TEXTO</i> ✍Fonte ✍Responsável ✍Categoria (de texto) ✍Original ✍Data ✍Título *Fonte: ENEM *Responsável: INEP/MEC *Categoria: Redação Escolar *Original: Manuscrito digitalizado *Data: 2002
FICHA 3 = <i>AUTORIA</i> ✍Idade ✍Origem - Região - Estado ✍Renda ✍Instituição	
FICHA 4 = <i>DESEMPENHO</i> ✍Nota COMP1 ✍Nota COMP2 ✍Nota COMP3 ✍Nota COMP4 ✍Nota COMP5 ✍Nota Global	

Nas fichas 1 e 2, conforme o esquema, vemos os itens cujo “valor default” é comum a todas as amostras. Os demais, embora dependentes do texto a que correspondem, são facilmente recuperados pela via automática.

Para os cabeçalhos também empregamos a tecnologia XML, o que tem sido recomendada pelos padrões internacionais e, de fato, usual nos dias de hoje. Dentre as vantagens do padrão XML para o cabeçalho está o fato de que ele permite a apresentação isolada do cabeçalho; i.e., podemos acessar um cabeçalho independentemente do texto a que se liga.

Face ao banco de dados que temos, que relaciona a cada um dos textos todos os traços do perfil, a inserção dos cabeçalhos no Corvo se deu de forma automática.

IV.1.1. 5 ANOTAÇÃO

Nosso leitor já está atento à empreitada desta pesquisa no campo da anotação de corpus. Neste momento da exposição o resultado que conseguimos alcançar em nosso método e materiais de pesquisa visando à anotação estarão em foco.

A etiquetagem de corpus que, no Corvo, se desenvolve nesta etapa de construção gera, pela primeira vez, uma nova versão dos textos. Isso porque, excetuando-se a anotação da hesitação, no plano da superfície textual, este é o momento em que a anotação lingüística é efetuada propriamente no corpus.

Tal como se anunciou de entrada, objetivou-se, neste trabalho, anotar um corpus em termos de *desvios da norma-padrão* detectados semi e automaticamente. Ou seja, buscamos marcar as ocorrências em que o redator não fez uso das regras prescritivas que padronizam os

textos escritos na variante culta do português⁴³. Ora, mas se parece não haver consenso sobre o que se poderia considerar desvio da norma-padrão do português, de que modo auferir esse título a alguma coisa dos textos? Ainda convivemos com referências discordantes entre si sobre quais seriam essas regras, e nos debatemos em meio a referências, de um lado, já obsoletas (p.ex., gramáticas como as de Napoleão de Almeida), de outro, muito recentes (p.ex.: manuais de redação de jornais de grande circulação no país). Decidimos, assim, lançar mão de um material de apoio; algo que tivesse levado em conta essa questão para indicar quais regras são prescritas com alguma concordância entre essas referências a que nos referíamos.

O objeto em questão é o *revisor gramatical automático ReGra*, já descrito aqui. No tocante às regras da norma-padrão, o uso do ReGra contribuiu para a definição de uma tipologia de desvio que, para nós, representou o ponto de partida para a tarefa de detecção dos desvios no corpus. Utilizamos a “tipologia de erros” do corpus PROBI, que reflete do tratamento do revisor ReGra sobre como rotular desvios.

Para a construção do corpus PROBI (MARTINS, 2002), os projetistas reuniram determinados fatos para os quais se exige o uso “correto” de recursos do português culto.⁴⁴ No ReGra, o que se toma como “erro” é algo para o qual o revisor deva estar preparado para agir. O conjunto desses fatos pode ser tomado como uma *tipologia de “erros”* (ou de “inadequação lingüística”) e encontra respaldo naquilo que se consideram referências da norma-padrão⁴⁵.

Partindo desse material, definimos aquilo que chamaremos de *tipologia de desvios* empregada no Corvo. As especificações são as seguintes.⁴⁶:

⁴³ Tal como supomos neste trabalho, essas regras foram apresentadas (ou deveria tê-lo sido) no período de escolarização até o nível do ensino médio.

⁴⁴ Falamos estritamente da revisão gramatical e de estilo, já que a revisão ortográfica automática, embora importante e complexa, não faz uso de tipologia de erros.

⁴⁵ São exemplos: as gramáticas tradicionais, alguns manuais didáticos de lingüística, manuais de redação do jornalismo, etc. Já discutimos essa questão nas seções teórica, mas importa lembrar, sobre o ReGra, de que para ele também houve decisões baseadas em corpus (na esfera do banco de dados lexical, p.ex.); a questão é: um corpus também pode influenciar na sua tipologia de erros? Prossigamos com a exposição.

⁴⁶ Aqui aproveitamos o modo como essas especificações aparecem no PROBI, inclusive com o exemplário e a proposição das etiquetas.

Tabela IV -10: Tipologia de desvio para anotação do Corvo

Tipo de desvios (etiqueta)	Classe de desvios	Descrição	Exemplos
cjc	USO DE CONJUNÇÕES	Indica inadequação no uso das conjunções e locuções conjuncionais	<p>☞ por causa que (porque)</p> <p>☞ mas porém (mas)</p>
cmt	CONCORDÂNCIA ENTRE MODOS E TEMPOS VERBAIS	Indica falta de concordância entre as formas dos verbos	<p>☞ se eu a ver (se eu a vir)</p> <p>☞ quero que ele vai (quero que ele vá)</p>
com	CONCORDÂNCIA NOMINAL	Indica falta de concordância entre os núcleos nominais e seus adjuntos ou predicativos	<p>☞ o menina (a menina)</p> <p>☞ casas bonita (casas bonitas)</p>
cop	COLOCAÇÃO PRONOMINAL	Indica inadequação no uso da próclise, ênclise e mesóclise	<p>☞ não comprou-a (não a comprou)</p> <p>☞ farei-a (fá-la-ei)</p>
cov	CONCORDÂNCIA VERBAL	Indica falta de concordância entre o sujeito e o verbo	<p>☞ as meninas foi (as meninas foram)</p> <p>☞ fazem dois anos (faz dois anos)</p>
cra	USO DE CRASE	Indica inadequação no uso do acento indicativo da crase	<p>☞ à partir (a partir)</p> <p>☞ à prazo (a prazo)</p>
det	USO DE ARTIGOS E DETERMINANTES	Indica inadequação no uso dos artigos e determinantes	<p>☞ todas meninas (todas as meninas)</p> <p>☞ cujo o pai (cujo pai)</p>
mal	USO DE MAU/MAL	Indica troca de “mal” por “mau” ou vice-versa	<p>☞ ele foi mau (mal) na prova</p> <p>☞ ele é muito mal (mau)</p>
ond	USO DE ONDE/AONDE	Indica troca de “onde” por “aonde” ou vice-versa	<p>☞ onde (aonde) você foi?</p> <p>☞ aonde (onde) estou?</p>
pre	USO DE PREPOSIÇÕES	Indica uso indevido de preposições e locuções prepositivas	<p>☞ após ao (após o)</p> <p>☞ o jeito dele (de ele) trabalhar</p>
pro	USO DE PRONOMES	Indica uso indevido de pronomes	<p>☞ entre ela e eu (mim)</p> <p>☞ eu vi ele (o vi)</p>
prq	USO DE POR QUE	Indica uso indevido de “porque”, “por que”, e “porquê”	<p>☞ não sei porquê (por que) ele saiu</p> <p>☞ porque (por que) você não vem?</p>
ptn	PONTUAÇÃO	Indica uso indevido dos sinais de pontuação	<p>o menino, saiu (o menino saiu)</p>
ptp	USO DE PARTICÍPIO	Indica uso indevido do particípio passado	<p>☞ ele foi pegado (pego)</p> <p>☞ ele tinha pago (pagado)</p>
reg	REGÊNCIA VERBAL	Indica inadequação no emprego da regência dos verbos	<p>☞ obedeceu o pai (ao pai)</p> <p>☞ assisto o filme (ao filme)</p>
ren	REGÊNCIA NOMINAL	Indica inadequação no emprego da regência dos substantivos, adjetivos e advérbios	<p>☞ ida no cinema (ao cinema)</p> <p>☞ bacharel de direito (em direito)</p>
ver	USO DOS VERBOS	Indica inadequação no uso das formas verbais	<p>a lâmpada acendeu (se acendeu)</p>

Tabela IV -10: Tipologia de desvio para anotação do Corvo (cont.)

Tipo de desvios (etiqueta)	Classe de desvios	Descrição	Exemplos
ses	SEM ESPECIFICAÇÃO	Inclui as ocorrências sem correspondência nesta tipologia	

Essa tipologia com **17** categorias (excluindo-se a etiqueta “ses”) certamente não exaure a variedade de desvios encontrados nos textos cuja população o nosso corpus representa. Contudo, a abrangência dessa tipologia é significativa, pois inclui os principais fenômenos de inadequação previstos nas gramáticas de referência do português, tais como a concordância, a crase, a pontuação e a regência.

Note-se que, embora a classe do desvio seja genérica, os exemplos tabulados demonstram o limite de cada uma. Por exemplo: o tratamento da pontuação enfoca os casos de emprego dos sinais na quebra de uma estrutura sintática padrão (sujeito e predicado) e tem em vista, especialmente, o uso da vírgula. Não é possível dizer, então, que a etiqueta incidirá sobre qualquer ocorrência de uso inadequado de qualquer sinal de pontuação. O que se destaca, enfim, é que essa tipologia não é inconsistente, embora se mostre incompleta; o que se vê, de fato, é que ela aponta fenômenos de desvios detectados correntemente no uso do português culto.

Mas dissemos atrás que a meta perseguida aqui foi a detecção e a anotação, mas que isso se daria pela via automática ou semiautomática. Mais uma vez, então, nos amparamos no revisor ReGra.

Para a detecção dos desvios no plano ortográfico, tratou-se de uma medida tomada em função do custo menor que representou para o nosso projeto: o ReGra detecta muito bem os problemas da ordem da ortografia e isso em tempo real, isto é, assim que o texto é digitado. No caso da detecção gramatical e de estilo, especialmente, além do fator custo, a questão do controle foi determinante. A detecção dos desvios pela via automática nos pareceu um modo satisfatório de controlar a subjetividade inerente ao processo de identificação dos desvios. Com o apoio do ReGra conseguimos uniformizar o tratamento dos textos, definindo, sob os mesmos critérios, quais de fato se consideraram ocorrências de desvios no corpus.

Determinados que estivemos em aproveitar ainda mais a participação do ReGra na anotação do nosso corpus, o passo seguinte foi modelamos uma ferramenta computacional, que

nomeamos como **Identificador do Desvio**, onde o revisor, que é o ReGra, desempenha papel fundamental. O aplicativo foi projetado para a minimizar o esforço do amador, oferecendo-lhe material de apoio de qualidade (uma tipologia de desvios, p.ex.). Detalharemos na sequência, a ferramenta em si bem como o modo como foi utilizada com Corvo.

Mas, antes, queremos adiantar ao leitor, que quanto ao percurso metodológico, a *detecção* dos desvios pela via desse aplicativo posteriormente se revelou insatisfatória, pouco convincente, e o percurso precisou ser reajustado. Pode-se dizer que no levantamento dos desvios do nosso corpus, esse aplicativo funcionou como apenas como auxiliar no processo, muito embora tenha respostas importantes a dar tanto para o próprio método de anotação como para a avaliação do revisor (seu desempenho, suas vantagens e limitações). O fato é, pois, que ele não foi a única estratégia do método empregado no Corvo.

Passemos adiante. Vamos conhecer a tal ferramenta e saber de sua interferência na nossa pesquisa.

IV.1.1.5.1 O Identificador do Desvio e a anotação automática: usuário e funcionalidades

No plano inicial, a tarefa de anotação contaria com um instrumental que nos permitisse:

- i - obter uma detecção o máximo eficiente e rápida;
- ii - apresentar cada uma das detecções de modo (computacionalmente) amigável;
- iii - validar a detecção;
- iv - aplicar a tipologia de desvio em uso no Corvo aos casos de detecção validados;
- v. validar a classificação do desvio;
- vi. controlar todas essas operações;
- vii. gerar versão anotada do texto em formato xml, com somente os casos de detecção e classificação do desvio validados.

Essas expectativas se realizaram, de fato, em todos os itens, mas não podemos dizer que a detecção dos desvios, como aparece em (i), tenha sido tão eficaz e, até em alguns casos, tão veloz quanto esperávamos. A partir disso, como veremos mais tarde, não o material ou o método como um todo, mas o núcleo duro desse nosso instrumental – ou seja, o revisor automático – frustrou a

estratégia metodológica empreendida. Afora essa questão, porém, podemos dizer que conseguimos cumprir o plano traçado para a ferramenta. Senão, vejamos.

Tomado, de partida, como um aplicativo para uso de especialista, alguém capaz de julgar e decidir sobre o processo de detecção e classificação dos desvios, procuramos construir um sistema com diversas funcionalidades e uma interface atraente. Isso se relaciona à meta em (ii) acima. A título de ilustração, apresentamos algumas telas do Identificador de Desvio (daqui por diante, IDv), em que essas funcionalidades, por exemplo, podem ser notadas:

Figura IV-14: Ilustração do IDv – apresentação e funcionalidades (I)

Como se vê pela ilustração, IDv é um programa preparado para receber qualquer arquivo texto. Seu processamento, no momento inicial, consiste na submissão do texto ao revisor gramatical automático – ReGra.

Nem todas as funcionalidades que se vê na interface estão habilitadas. Mas pode-se observar que uma classe de usuários nova está em perspectiva em IDv: a de avaliadores de sistemas de revisão. Consideraremos essa questão; antes, convém destacar um detalhe importante na ilustração acima.

Na imagem, se vê aquilo que chamamos de *questões de controle*, que são colocadas ao usuário, impondo-lhes restrições. Da perspectiva do público final da análise de corpus, essas restrições se refletem na confiabilidade desse sujeito no aplicativo; da perspectiva do projetista isso implica evitar que determinados equívocos sejam cometidos desnecessariamente,

comprometendo a confiança no resultado final. Em diversas etapas do IDv essas questões de controle são disparadas. Elas devem ser distinguidas, porém, das “questões de validação”, mas apenas nessa relação com a funcionalidade de IDv. Adiante mostraremos o processo de validação. O caso é que a funcionalidade das questões de controle visam a uma classe de usuário de IDv, ao passo que as questões de validação é uma funcionalidade visando à nova classe de que falávamos atrás. Tomemos de perto, então, esse elemento incluído no curso da criação do aplicativo, que é o usuário avaliador. Introduzido esse aspecto, poderemos continuar com o processo de anotação do Corvo, progredindo na utilização do aplicativo.

Conforme abordamos na seção teórica, um sistema de revisão automática pode ser avaliado a partir dos chamados *corpus de teste*, e um primeiro passo disso é o processamento propriamente dito do texto pelo revisor. Pode-se, a partir de então, avaliar o desempenho do sistema, julgando, por exemplo, se a sua intervenção foi competente ou não. Para o usuário linguista de corpus que, como nós, objetivamos a detecção e classificação do desvio, o que importa é rejeitar ou manter uma detecção verdadeira com classificação consistente a ela relacionada. Para um usuário avaliador do sistema de revisão, observa-se a detecção na expectativa de separar as detecções devidas ou não e as omitidas. Ou seja, não temos aqui uma detecção e classificação do desvio simplesmente, mas também uma rotulação da ocorrência como um fator positivo ou negativo do desempenho do revisor.

Assumindo o interesse em auxiliar esse usuário avaliador, introduzimos em IDv um conjunto de operações, e passamos a trabalhar com o Corvo também com essa perspectiva de observar o comportamento do ReGra.

Em termos de funcionalidade, por exemplo, essa nova abrangência de IDv podem ser observadas na ilustração abaixo, em que trazemos o momento de validação do desvio. Veja-se que existe um texto em que alguns desvios foram detectados. Eles são mostrados ao usuário em forma de *links*. A classificação da intervenção e do desvio são disparadas em sequência.

A validação consiste em indicar: a) se a ocorrência é um verdadeiro positivo, i.e., se o revisor interveio corretamente; e b) qual a categoria genérica do desvio – se é um problema mecânico, ou ortográfico, ou gramatical ou de estilo. Vejamos a ilustração.

Figura IV-15: Ilustração de IDv: apresentação e funcionalidades (II)

Aqui temos uma daquelas questões colocadas ao usuário. Desta vez aparece uma *questão de validação*, ou seja, supervisiona-se o trabalho do usuário no sentido de “validar” as classificações a serem efetivadas.

A essa altura é importante mencionar que a edição do texto em processamento não está autorizada ao usuário. Mas ele é chamado à proposição de diversas informações no decorrer do trabalho com IDv. É o que podemos verificar na próxima ilustração. Note-se que é uma função disparada quando a detecção é considerada um falso negativo, i.e., um momento em que o revisor interveio indevidamente.

Validação dos Desvios

O direito de votar, a muito tempo não eram todas as pessoas que podiam exercer o voto, apenas os mais ricos, donos de terras, mas nós lutamos e pouco a pouco vamos concedendo este direito a todos, até chegarmos a época de hoje, em que o voto não é apenas um direito e sim uma obrigação. Será que esta luta valeu a pena?

Input

Entre com alguma observação/comentário: (Opcional)

OK Cancel

Desvio:

Concordância Nominal

Mais Informações:

As mulher saiu. (Inadequado)
 A mulher saiu. (Adequado)
 Dois crianças caíram. (Inadequado)
 Duas crianças caíram. (Adequado)

O artigo, o pronome, o numeral e o jetivo determinantes devem sempre concordar em gênero e número com substantivos a que se referem.

validação

☐ verdadeiro ☒ falso

classificações

☐ mecânico ☐ ortografia
☒ gramática ☐ estilo

Finalizar **Cancelar**

Na verdade, a *caixa de diálogos* do sistema de revisão com o qual lidamos parece nos dar o tom dos comentários, a maneira de padronizarmos esse texto de preenchimento. Para nós, no ReGra, essas caixas de diálogo são especiais em dois momentos: o primeiro, em que ele apresenta a sugestão de alteração, e o segundo, em que ele apresenta o chamado “Mais Informações”. Essas duas caixas foram trazidas para o IDv e são carregadas ao tempo do clique sobre o link do desvio detectado, tal com se vê à direita na ilustração acima.

96

Com a validação concluída (na interface, o clique sobre “finaliza”), o que IDv produz é um texto, que podemos considerar uma versão “intermediária” da amostra anotada do corpus. Isso porque a anotação do desvio não se completa neste estágio; porém, já é uma nova versão do corpus, visto que temos aqui um texto em XML.

Vale dizer que, quanto ao Corvo, os textos de *input* da etapa de validação são aqueles que formatamos com o processo de compilação. Ou seja, são textos já com etiquetas de superfície. Assim, nessa nova versão, além do novo formato em XML, o que temos é o acréscimo (e não a inclusão) de etiquetas.

Da perspectiva do avaliador, os resultados (parciais) desse texto intermediário são de grande valor. Vamos discorrer sobre essa questão, aproveitando para apresentar novas telas do IDv e apontar outras de suas funcionalidades.

O leitor atento deve ter notado, na barra de *menu*, as opções “Relatórios”⁴⁷. O objetivo aqui é gerar relatórios descritivos do processamento do revisor sobre o texto. Planejamos esses relatórios com algumas especificidades, mas antes de qualquer coisa já se pode notar que esse tipo de funcionalidade tem sob perspectiva um usuário que quer “documentar” aquilo que fez até então com o aplicativo. Os relatórios são importantes também para fornecer dados para tabulação sobre o desempenho do revisor, para coletar as informações adicionais (p.ex.: comentários), e tantos outros interesses. De todo modo, os relatórios representam para o avaliador do revisor um material informativo e estruturado.

Foram planejados dois tipos de relatórios e duas formas de acessá-los. Chamamos de *relatório simples* aquele que corresponde a apenas um texto; e de *relatório síntese* aquele que reúne os dados de mais de um texto. Os relatórios sínteses são criados desde que existam os relatórios individuais. Ambos os tipos de relatórios também são documentos XML e apontam: a) data e hora do processamento; b) a versão do revisor utilizada no processamento; c) quantos e quais os arquivos (textos) a que se correspondem; d) resultado da validação, i.e., quantos e de que tipo são os desvios aferidos; e) informações adicionais sobre o desvio fornecidas pelo revisor (Sugestão de alteração e Mais informações); f) contexto de ocorrência, com até sete janelas à esquerda e à direita; g) comentários, críticas e sugestões editadas pelo usuário (quando houver).

⁴⁷ Também há as opções “Arquivo” e “Editor”. A primeira é sobre o que vimos tratando até aqui. A segunda será detalhada na sequência.

A fim de que o usuário pudesse utilizar os relatórios para a observação mais estreita do corpus, dos dados ou do contexto de ocorrência, por exemplo, julgamos que o modo de apresentação dos relatórios deveria ser o mais dinâmico possível. Por isso, também investimos na *visualização* dos relatórios, criando outra funcionalidade do aplicativo. O conteúdo é aquele que indicamos acima, mas, no modo de visualização, o usuário pode “navegar” pelo relatório, acessando os *links* disponíveis, os quais disparam janelas separadas contendo as informações pertinentes.

As imagens abaixo ilustram algumas telas de IDv na função Relatórios. Nas últimas, mostramos os relatórios no modo visualização, de onde é possível notar algumas possibilidades de navegação do hipertexto, conforme acabamos de mencionar:

Figura IV-17: Ilustração de IDv: apresentação e funcionalidades (IV)

Visualiza Relatório Simples

Principal

RS_C2A_1-1.xml

RS_C2A_3.xml

Relatório gerado em: 22/4/2007 as 14:45:48

Utilizando a versão 12.0 do ReGra

Contendo 1 redação

Redação pertencente:

C2A_3

Desvios aferidos:

Nenhum desvio foi aferido.

Visualiza Relatório Síntese

Principal

para ilustracao 1.xml

Relatório gerado em: 22/04/2012 10:00:00

Utilizando a versão 12.0

Contendo 36 redações

Redações pertencentes:

[C3C_1](#)

[C3C_10](#)

[C3C_11](#)

[C3C_12](#)

[C3C_13](#)

[C3C_14](#)

[C3C_15](#)

[C3C_16](#)

[C3C_17](#)

[C3C_18](#)

[C3C_19](#)

[C3C_2](#)

[C3C_20](#)

[C3C_21](#)

[C3C_22](#)

[C3C_23](#)

[C3C_24](#)

Redação

O Direito de Votar

É muito importante votar, só assim vamos conseguir mudar o nosso país. É importante ter paciência e ser detetive para saber um pouco de cada candidato porque não adianta votar depois que esses já estiverem lá. Para que isso não aconteça temos que votar com cuidado e não se vender por qualquer coisa que seja a quantia.

É com muita vontade que vamos votar e votar bem. Me orgulho de ter o meu país. Quando for a hora eu estarei lá para mudar o Brasil.

So nós vamos poder mudar nossas vidas economica, saúde e paz em nosso país.

Cabeçalho

Sair

Figura IV-20: Ilustração de IDv: apresentação e funcionalidades (VII)

É mister dizer que nem todas as funcionalidades programadas para IDv estão habilitadas. Aliás, esse aplicativo ainda carece de refinamentos importantes, que não foram possíveis no âmbito deste trabalho. Uma das funções desabilitadas é, justamente, o cabeçalho.

Se notarmos a tela de abertura de IDv, veremos o ícone *Cabeçalho* no campo reservado para as ferramentas de processamento textual. A idéia era oferecer ao usuário a possibilidade não só de inserir o cabeçalho dos textos submetidos à anotação, mas, principalmente, de permitir a visualização dos cabeçalhos já existentes, como é o caso do Corvo, em que a documentação dos textos foi promovida antes da anotação.

Essa mesma intenção justifica a presença do cabeçalho na tela de visualização do texto, na dinâmica dos relatórios. Uma vez que o usuário está navegando pelo relatório e quis acessar determinado texto, seria interessante disponibilizar os dados de autoria, por exemplo, a fim de ampliar a visão crítica do pesquisador. Por restrições de difícil solução, porém, não temos essa funcionalidade disponível hoje em IDv. Voltaremos a essa questão em nossas considerações finais.

Como última função de IDv o que se tem é um novo processamento do texto, aliás, a etapa final da anotação (do desvio). Trata-se do *Editor de Desvio*, que o nosso leitor já deve ter notado na interface.

Na verdade, esse Editor fora planejado para funcionar separadamente; ou seja, queríamos dois aplicativos, um para validação das detecções e classificação conseguidas com o ReGra, e outro que fosse livre, independente de ferramentas de apoio, para uso do especialista na edição da anotação. No entanto, tivemos de compactar os aplicativos, habilitando, ambos, numa mesma interface e mais: tornando a edição dos desvios um processo dependente da etapa anterior, que é a validação.

O usuário, finalizando a validação e gerando ou não os relatórios, completa a anotação dos desvios com o auxílio do Editor. O texto de *input*, assim, é um documento XML, com etiquetas de superfície textual e etiquetas dos desvios validados.

Como foi possível observar, o uso do ReGra, até aqui, nos devolveu um levantamento dos verdadeiros positivos e dos falsos negativos, i.e., as detecções verdadeiras (o revisor interveio devidamente) e as detecções falsas (o revisor não interveio devidamente). Mas o que dizer dos desvios não detecções pelo revisor? Os casos, que costumamos chamar de *omissões*, são de fundamental importância. Há espécies de desvios para os quais o revisor não está preparado, boa parte das vezes em virtude da complexidade lingüística que imputa ao processamento automático do texto, como vimos na exposição teórica. Outros momentos há em que algum “ruído” presente na sentença dificulta o “casamento” das seqüências, o que leva o revisor a tomadas de decisão que não favorecem a detecção do desvio, o que aconteceria se não fosse o ingrediente não esperado (p.ex.. mal emprego dos sinais de pontuação, que prejudica a capacidade de estruturação sintática de uma sentença pelo analisador sintático do revisor).

Na prática, marcar as omissões do revisor significa interpretar o texto do mesmo modo que o revisor faria; é buscar “imitar” o modo como ele compreende e analisa o texto, e dialoga com seu usuário. As omissões são os casos em que o revisor deveria intervir, mas não interveio; é um falso positivo. Um comportamento indesejável do seu usuário, já que, julgando ser auxiliado por um instrumental para esse fim, ele não é alertado de que produziu um desvio.

A tarefa do especialista, na edição dos desvios, sua classificação, etc., no caso do Corvo (e cremos que para outros desse tipo), não prescindiu de um material padronizado, criterioso, em

que pudesse basear suas intervenções. Partimos da tipologia de desvios, já empregada na etapa de validação, para nos apoiar, inclusive na classificação do desvio (se desvio de pontuação, de regência, de regência nominal ou verbal, e assim por diante).

Aqui, sim, podemos dizer que a etiqueta “ses” (sem especificação) for largamente empregada, já também observamos – e marcamos – casos que, supostamente, estão fora da tipologia de base (a tipologia do PROBI/ReGra). Não só por investirmos nisso, mas, certamente, em grande medida por conta de, é que se decidiu, no projeto, que nem todas as omissões seriam marcadas, ainda que detectadas pelo especialista. Tratou-se de uma medida em favor do que era possível fazer com qualidade num trabalho como este em que o corpus é de estudo. Abordamos esse tema das omissões e o que ela implicou no curso da nossa investigação no capítulo da análise crítica, na sequência.

A edição dos desvios, auxiliada, então, pelo Editor em IDv, pode ser visualizada nas telas seguintes:

Figura IV-21: Ilustração de IDv: apresentação e funcionalidades (VIII)

Figura IV-22: Ilustração de IDv: apresentação e funcionalidades (IX)

Em termos de *layout* e opções de navegação, o Editor de desvios (daqui por diante, EDv) se mantém nos padrões do aplicativo. O leitor já pode notar a funcionalidade dos relatórios também aqui presentes, indicando-nos que esse processo de edição é documentado aos moldes do que descrevemos anteriormente. Verifica também o um elemento destacado no texto de entrada (“*.pois*”), que é, justamente, o desvio validado. É que, como dissemos, para esse texto descartou-se os falsos negativos, mas os verdadeiros positivos foram mantidos.

A caixa de diálogo que se vê na interface (Fig. 22) é a que se dispara quando determinada ocorrência é selecionada para anotação. Pode-se verificar com ela que, para se concluir a anotação do desvio, alguns dados deverão ser preenchidos⁴⁸. Exceto a janela para os comentários, que é opcional, as demais devem necessariamente trazer: a) um texto com o que se poderia considerar um tópico do Mais Informações do revisor; e b) uma sugestão de mudança, conforme também se propõe no revisor. Com o Corvo, foi definido que as frases do Mais Informações teriam similaridade com as mensagens explicativas do ReGra, e que as sugestões de mudança somente apareceriam, se pudessem levar o redator à alteração efetiva, i.e., alterar sem ser preciso compreender a regra de correção.

⁴⁸ Veja-se que esse caso se comporta como uma questão de validação do tipo obrigatório; ou seja, para ser rotulado como desvio, é preciso fornecer as informações solicitadas e, só partir disso, prosseguir com o aplicativo.

Ressalta-se, por fim, a classificação do desvio nas categorias previstas pelo revisor (se desvio gramatical, ortográfico, etc.). É preciso dizer, contudo, que no nosso caso, já que optamos pela observação de determinados casos de omissão e não todos, temos apenas desvios do tipo gramatical anotados no corpus.

Aliás, a este ponto retomamos a perspectiva do Corvo, especialmente. Voltemos à descrição de seu percurso metodológico, posto que, a par do que acabamos de expor, nosso leitor agora sabe que o Corvo tem uma versão anotada dos textos, rotulando-se os casos de desvios. Temos, além disso, uma etiquetagem que serve à avaliação do desempenho do revisor ReGra.

Ocorre que, embora IDv tenha sido bastante eficaz em alguns aspectos, o processamento não ocorreu no corpus inteiro. Em termos exatos, a versão do Corvo em formato XML é composta de **157** textos. O aplicativo não funcionou para 92 textos do corpus, que permanecem na versão base, a propósito, uma versão anotada, mas quanto à superfície textual.

IV.1.1.5.2 A detecção de desvio: formação de indicadores e levantamento de dados

Em vista dos objetivos da investigação, em especial, o de saber quais são os desvios e com qual frequência ocorrem no nosso corpus, a detecção dos desvios precisou ser realizada de outro modo e desta vez no corpus inteiro. Podemos dizer que através desse novo trajeto metodológico faltou para a anotação a colocação das etiquetas nas amostras e a detecção das omissões (ou falsos positivos), visto que, no mais, promoveu-se por essa estratégia o essencial exigido na anotação, no nosso caso, a detecção e a classificação (em desvio e em comportamento do revisor) das ocorrências.

Mas não fora somente o problema operacional de IDv, não processando certo número de textos, que determinou um novo caminho. Na verdade, a capacidade de detecção de IDv mostrou-se insatisfatória desde muito cedo, mas apenas agora trazemos essa questão. O que houve foi uma impossibilidade de trazer o ReGra por completo para o aplicativo, e isso por restrições alheias a este trabalho. Como sabemos, o revisor é um produto patenteado e não tivemos acesso a algumas de suas funcionalidades. A principal delas foi o módulo ortográfico, responsável pela detecção, correção e aconselhamento de desvios desse tipo (grafia, acentuação, etc.). Em outras palavras,

IDv não aplicou o ReGra e todo o seu potencial aos textos; a ortografia, por exemplo, não foi tratada.

E não trazer o módulo ortográfico gerou outros incômodos. O processamento sintático dos textos certamente é afetado no ReGra; ora, como analisar corretamente a estrutura sintática de uma sentença, assumindo como correta a ortografia do texto? Os resultados da detecção gramatical, por isso, foram tomados com desconfiança. Em princípio, por exemplo, seria muito maior o número de omissões da ferramenta, ainda que esse levantamento não fosse pretendido e, de fato, não se realizou com o aplicativo. Ou seja, era muito mais provável que o desempenho do ReGra fosse ruim em termos de falsos positivos, uma vez que a análise ortográfica do texto não aconteceu.

Mas o fato é que, a fim de termos parâmetros mais consistentes para a análise, nossos textos precisariam ser submetidos a uma avaliação da ortografia. A saída encontrada, então, foi submeter novamente o corpus à detecção dos desvios, sem avançar para a anotação, isto é, identificar e registrar as ocorrências, mas não inserir as etiquetas junto às ocorrências no texto. Para tanto, trabalhamos na plataforma Word para abrir os textos do corpus e passamos a registrar o levantamento dos dados numa planilha de dados.

Como já se sabe, o ReGra está embutido no sistema de revisão do Word^{MS}. Submetemos uma a uma as amostras do Corvo e as detecções, incluindo-se as ortográficas, foram mostradas de forma visual; como se sabe, destaques em vermelho para os desvios ortográficos, destaques em verde para desvios por fatores mecânicos, desvios de estilo e os desvios gramaticais. Na imagem a seguir, vemos o momento da análise do texto e o preenchimento da planilha:

Figura IV-23: Processo de detecção dos desvios e levantamento de dados

Desta vez, a totalidade dos 249 textos foram processados pelo revisor. Podemos dizer que o levantamento só não foi mais exaustivo sobre os desvios, porque não especificou o tipo, distinguindo-os apenas como gramaticais ou ortográficos, mecânicos ou de estilo.

Ao realizarmos a avaliação das detecções, propusemo-nos também a separar as ocorrências pelas categorias de desempenho do revisor. Criamos alguns *indicadores de desempenho*, que refletem a classificação empregada no ReGra: verdadeiros e falsos positivos ou negativos. No levantamento produzido, foram observados e contabilizados: a) os verdadeiros positivos em cada um dos níveis de análise: mecânico, ortográfico, gramatical e de estilo; b) falsos negativos somente da análise gramatical e de estilo; c) nenhum falso positivo.

Com esse trabalho concluído, chegamos ao fechamento deste corpus de estudo que é o Corvo. Mostramos, no decorrer desta seção, de que modo não só o corpus foi construído e se mostrou adequado à análise, como também nossas opções metodológicas: o apoio em recursos lingüístico-computacionais de intersecção, como é o ReGra, a tipologia de erros do PROBI, nossos próprios aplicativos – IDv e EDv . Indicamos, além disso, o referencial que será usado na análise dos dados, que são as tabelas estatísticas, os relatórios de desempenho, o levantamento semi-automático das detecções.

V. Análise do Corvo: reflexões e propostas

O Corvo é um corpus para estudos dos mais diversos tipos, pois traz uma espécie de texto rica, que seguramente interessam aos pesquisadores do português. Nosso olhar sobre os textos é do ângulo de quem se preocupa tanto com os fenômenos lingüísticos ali presentes quanto com os indivíduos que os produziram. Vemos pessoas, que se expressaram por escrito em língua materna; que estiveram sob as condições peculiares de uma avaliação dessa sua escrita, e o texto que temos é, desse modo, um registro lingüístico de grande valor. Quisemos tomar um extrato dos textos produzidos no contexto do ENEM, transformá-lo num conjunto formatado de dados, e chega o momento, enfim, de apontarmos algumas particularidades observadas no corpus, no cruzamento dele com a base de dados do perfil, no processamento dos textos pelo revisor.

Como vimos, investigamos ocorrências de desvio da norma padrão no nosso corpus, procurando detectar tais ocorrências por meio de uma tipologia de desvio previamente definida e um mecanismo computacional de auxílio. Além disso, procuramos integrar os dados textuais com os dados extra-textos, criando indicadores relevantes quer para descrever o sujeito redator quer para avaliar o instrumental de apoio na geração dos dados do texto, i.e., o revisor automático ReGra.

Tomemos o ReGra, então, como ponto de partida. De modo geral, o auxílio do revisor foi significativo, mas gostaríamos de tê-lo aproveitado mais, sobretudo funcionando para tratar melhor esse tipo de texto com franca presença de desvios. Mas, como estamos a afirmar sobre funcionamento melhor e franca presença de desvios? De onde vem essa confiança? – nos perguntariam.

Mesmo com um desempenho irregular, provocaríamos esse nosso interlocutor, dizendo que o ReGra interveio em quase todos os textos, mostrando que, apesar de tudo, os textos de fato carecem de revisão da escrita, e fundamentalmente em nível lexical, ortográfico, o do ritmo das frases, com a importante questão da pontuação, e assim por diante.

Vê-se, pois, que o quadro diante do qual estamos deve ser abordado com alguma estratégia, afim de que o nosso leitor não tome essas afirmações como certezas e muito

apressadas. Daqui por diante, mostraremos os resultados da nossa pesquisa, analisados sob diversas perspectivas, cada qual em razão de um plano de observação em que vimos nos colocando ao longo da exposição: desde a de projetista e “usuário” de corpus à de projetista e usuário do ReGra; a de professora de sujeitos como os que se observaram neste trabalho à de avaliadora desses mesmos indivíduos numa banca do ENEM.

V. 1 As intervenções do ReGra e o perfil do corpus no plano ortográfico

Partindo do levantamento realizado com o ReGra na plataforma Word, em que aparecem as intervenções ortográficas inclusive, constatamos que o revisor interveio em 244 textos, ou seja, em **98%** do corpus. Vemos, assim, que se trata de textos com problemas reais de adequação lingüística. Tanto mais se observarmos o quadro de variação abaixo:

Figura V-1: Comportamento do ReGra sobre o Corvo: quantidade de intervenções

Como é possível observar, não se trata de intervenções esporádicas do revisor. Em **75%** das vezes, o revisor interveio mais de uma vez, chegando a intervir quatro vezes num único texto em 10% dos casos. Ora, considerando-se as limitações do revisor, que não age tal qual o revisor humano, podemos dizer que esses textos estão, de fato, entre aqueles de situação crítica no uso dos padrões da variante culta do português. Mas devemos considerar melhor esse desempenho do revisor.

As intervenções do ReGra foram separadas em:

☞ **Desvios Mecânicos:** são *verdadeiros positivos*. As detecções envolvem os casos de: repetição de palavras; emprego de maiúscula em início de sentença; ausência de pontuação em fim de sentença; dupla pontuação, dentre outros.

☞ **Desvios Ortográficos:** são *verdadeiros positivos*. Aqui, capturam-se os casos de: erro de grafia; acentuação gráfica ausente ou inadequada; emprego de inicial maiúscula em nomes próprios. Porque reconhece itens lexicais a partir de seu dicionário, a revisão ortográfica também acaba detectando os casos de oralidade (p.ex.: “tá”), os diminutivos (p.ex.: “propostinhas”) e os neologismos (p.ex.: “marginalismo”), todos considerados erros ortográficos pelo revisor.

☞ **Desvios Gramaticais VP:** são *verdadeiros positivos*. Nas detecções aparecem as inadequações listadas na tipologia de desvios; ou seja, são inadequações no emprego da concordância (nominal ou verbal), no uso de pronomes, conjunções e tempos verbais, na aplicação da crase, etc.

☞ **Desvios Gramaticais FN:** são *falsos negativos*. A tipologia de desvios também é aplicada aqui, nos possibilitando detectar as intervenções indevidas do revisor no âmbito gramatical.

☞ **Desvios de Estilo:** são *verdadeiros positivos*. Verificam-se detecções relacionadas à extensão dos períodos; à repetição de palavras em sequência (p.ex.: o chamado “queísmo” em que se tem períodos longos, com sentenças iniciadas invariavelmente pelo “que”); à grafia de números (p.ex.: aconselha-se a grafia por extenso dos números até 10), etc.

Como se vê, com exceção da revisão gramatical, todas as intervenções do ReGra aqui consideradas são intervenções devidas, corretas, que deveriam acontecer. Observe-se, então, qual o comportamento do revisor no corpus, quando separamos as intervenções segundo essa categorização acima:

Figura V-2: Distribuição das intervenções do ReGra

Esse quadro nos revela fatos bastante significativos. Veja-se, primeiramente, que a ortografia é uma classe de intervenção com maior incidência, e que o contrário acontece com a revisão ligada ao estilo. Notadamente, portanto, os textos são precários no trato da ortografia, sendo esse o principal fator de inadequação à variante culta. São 214 intervenções devidas do

revisor, ou seja, em **86%** dos textos, houve casos de desvios ortográficos corretamente identificados.

Vê-se, em seguida, que as intervenções no plano gramatical não apresentam um quadro tão satisfatório, já que, se por um lado o revisor detectou desvios gramaticais verdadeiros em 59% dos textos, por outro, são 55% de textos com intervenções indevidas nesse mesmo quesito. O revisor, grosso modo, acerta tanto quanto erra, o que nos remete à idéia de que, se estamos no universo de textos com alta incidência de desvios ortográficos, é bastante provável que isso interfira no desempenho da revisão gramatical da ferramenta, elevando os casos de falsos negativos. Tanto mais se notarmos o percentual de detecções mecânicas, que é alto (25% dos textos apresentam desvios desse tipo), o que seguramente compromete a análise estrutural da sentença pelo revisor.

Os textos do Corvo são de fato precários na ortografia e podemos até dizer que, se a ferramenta pudesse funcionar com maior propriedade, detectaria um número muito maior de desvios. A acentuação, por exemplo, foi fortemente negligenciada pelos redatores. A palavra “país”, por exemplo, de alta frequência no corpus, em virtude do tema da redação, não foi acentuada muitas vezes, o que, por sua vez, acarretou falsos negativos gramaticais em segmentos como “nosso país”, em que se pede a concordância. Mas esses casos foram detectados em sua totalidade pelo revisor; o problema está na não detecção de desvios de acentuação também frequentes (ausência do acento em “ê” e “está”, p.ex.) e que também deveriam levar o revisor a erros.

Mas a questão da ortografia afeta realmente o desempenho do revisor no plano gramatical, e pudemos comprovar isso com o nosso aplicativo.

Como se sabe, o corpus foi em grande parte processado pelo IDv, para a identificação e classificação dos desvios visando à anotação. Mas o revisor nele embutido não contou com o módulo de revisão ortográfica, de maneira que a revisão gramatical se deu sem se reconhecer os problemas ortográficos presentes nos textos. Ora, comparando-se o desempenho do ReGra no processamento via Word e via IDv, verificamos grande diferença e em favor do primeiro. Vejamos os dados:

Tabela V-1: Comparação de desempenho do ReGra (I)

Detecções	ReGra no Word*	ReGra no IDv*
Nenhuma detecção	5	89
Desvios Gramaticais VP	147	43
Desvios Gramaticais FN	136	31

*Quantidade de textos

Apesar da evidente vantagem do uso do ReGra em sua versão completa, i.e., com o módulo ortográfico funcionando, julgamos que talvez esses dados estivessem viciados, já que, como sabemos, o IDv não processou um número considerável de textos (exatos 89). Então separamos todos os textos processados pelo IDv e comparamos com o tratamento que tiveram quando do processamento no Word. Vejamos o resultado:

Tabela V-2: Comparação de desempenho do ReGra (II)

Detecções	ReGra no Word*	ReGra no IDv	No. de textos
Desvios Gramaticais VP	52	51	43
Desvios Gramaticais FN	31	36	31

*Quantidade de ocorrências

Diante desses dados, podemos constatar que houve pequeno aumento dos falsos negativos no processamento via IDv, mas, de modo geral, o comportamento do revisor manteve-se regular nos dois ambientes. É que, na verdade, os números absolutos não revelavam o detalhe de que alguns desses textos (14, para sermos exatos) não sofreram intervenção do ReGra no ambiente Word. Levando-se isso em conta, poderíamos até chegar a dizer que o revisor, no IDv, se superou, detectando ocorrências devidas de desvios gramaticais não capturadas no Word. Mas vejamos como está o quadro comparativo, quando tomamos o conjunto de textos para os quais o revisor, no IDv, não interveio de forma alguma.

Tabela V-3: Comparação de desempenho do ReGra (III)

Detecções	ReGra no Word*	ReGra no IDv*	No. de textos
Desvios Gramaticais VP	56	0	89
Desvios Gramaticais FN	63	0	89

*Quantidade de ocorrências

Desta vez o panorama é bastante elucidativo. Se o ReGra, via Word, não interveio devidamente em 14 textos, via IDv a não intervenção aconteceu num número muito maior de

textos. Além disso, são 56 ocorrências verdadeiras de desvios gramaticais não detectadas no IDv, mas, o que se destaca é o número considerável de falsos negativos do ReGra na plataforma Word, mostrando que, nesse conjunto de textos, o desempenho do ReGra foi insatisfatório, já que errou mais do que acertou.

Tomemos, agora, o *ranking* de textos ordenados pela maior frequência de desvios ortográficos. Observem as demais categorias de intervenção assinaladas para cada texto e será possível verificar que quanto mais ocorrências de desvios ortográficos aparecem no texto, menor é a capacidade de intervenção devida do revisor no plano gramatical.

Tabela V-4: Distribuição das intervenções do ReGra

Arquivo	Desvios Ortográficos (VP)	Desvios Gramaticais (VP)	Desvios Gramaticais (FN)	Desvios Mecânicos (VP)	Desvios de Estilo (VP)
C3A_4	22	0	0	0	0
C3C_73	15	2	0	0	0
C1B_2	14	0	0	0	0
C4B_3	14	1	2	1	0
C3C_23	14	0	1	1	0
C3B_10	13	0	0	0	0
C2B_7	12	0	0	0	0
C2C_34	12	1	1	1	0
C2C_52	12	1	2	1	0
C2C_7	12	0	1	2	0
C3C_72	12	0	1	18	0
C4C_2	12	0	0	0	0
C5C_11	12	1	2	0	0
C3B_11	11	0	0	0	0
C2C_19	11	1	1	0	0
C2C_42	11	0	0	0	0
C2C_56	11	1	2	0	0
C2A_2	10	0	0	0	0
C1C_2	10	1	1	1	0
C2C_11	10	0	0	0	1
C2C_32	10	0	2	0	0
C2C_50	10	1	1	2	0
C2C_41	9	0	0	1	0
C2C_46	9	0	1	2	0
C2C_49	9	0	2	1	0
C3C_65	9	1	2	7	0
C3C_75	9	1	3	0	0
C5C_6	9	4	2	2	0
C3A_1	8	0	0	0	0
C1B_3	8	4	0	2	0
C1B_4	8	0	0	0	0

Tabela V -4: Distribuição das intervenções do ReGra (cont.)

Arquivo	Desvios Ortográficos (VP)	Desvios Gramaticais (VP)	Desvios Gramaticais (FN)	Desvios Mecânicos (VP)	Desvios de Estilo (VP)
C2C_12	8	0	2	0	0
C2C_44	8	0	2	0	1
C2C_51	8	3	1	2	0

*Quantidade de ocorrências

Pode-se verificar, assim, que à medida que o texto apresenta maior corretude no plano ortográfico, no plano gramatical não só as intervenções devidas acontecem com mais frequência como se reduzem os casos de intervenções indevidas. Dito de outro modo, é maior a chance de o revisor errar em textos com mais incidência de desvios ortográficos ou de não intervir de forma alguma nesses casos.

Por ora, essas informações são o bastante para caracterizarmos minimamente o nosso corpus. Abordaremos, a seguir, as detecções gramaticais e, então, poderemos retomar essas especificidades da ortografia com base em novos dados.

V. 2 As detecções gramaticais e a tipologia de desvios

Algo que sempre insistiremos neste trabalho é que os textos com os quais nos ocupamos são insatisfatórios na adequação à norma padrão. Não se espera, então, ausência ou rara presença de desvios de toda sorte; ao contrário, tratar a quantidade e a variedade deles é que se torna o nosso desafio.

O Corvo apresentou um quadro de detecções dos desvios gramaticais tal como se vê abaixo:

Tabela V-5: Quadro geral das detecções dos desvios gramaticais

Detecções		Nenhuma ocorrência*	Uma ocorrência*	Mais de uma ocorrência*	Total (de textos)
Word	Desvios Gramaticais VP	102	83	64	249
	Desvios Gramaticais FN	117	85	47	
IDv	Desvios Gramaticais VP	115	37	8	160
	Desvios Gramaticais FN	129	26	5	

*Sobre quantidade de textos

Já sabemos que as detecções gramaticais não aconteceram num número elevado de textos do corpus, em grande medida pelos problemas de natureza ortográfica, que são frequentes no Corvo e interferem sobremaneira no processamento sintático do revisor em que nos apoiamos para o levantamento das detecções. Mas importa ressaltar, desta vez, que, ainda assim, mais da metade do corpus apresenta detecções validadas, i.e., há desvios gramaticais confirmados. Como se vê no quadro acima, na detecção via Word, desvios verdadeiros estão presentes em 147 textos (ou 59% do corpus), muitos dos quais com mais de uma ocorrência.

Como nosso leitor sabe, não foi fácil empreender a tarefa não só de detectar, mas sobretudo de classificar os desvios gramaticais, uma vez que somos críticos do posicionamento ortodoxo das gramáticas tradicionais, que muitas vezes rejeitam a interferência do uso da língua, especialmente quando se trata de esvaziar determinadas regras normativas. Apesar disso, assumimos esse desafio, já que nosso contexto é a variante culta e nela preside o tal elemento chamado norma padrão.

Na falta de um referencial de consenso sobre qual seria essa norma padrão, definimos um conjunto de tópicos gramaticais como nosso ponto de apoio, inspirados no material empregado no funcionamento do ReGra. Denominamos esse material de *tipologia de desvios*, e nela incluímos 17 categorias de desvio gramatical supostamente tomadas como essência da norma padrão, com prescrições que se aplicam em textos produzidos por um “usuário” mediano do português culto. Restou-nos o compromisso de avaliarmos a eficácia dessa tipologia num corpus como o Corvo, de aprendizes da variante culta e de textos muito pouco aplicados no uso da norma padrão.

Uma parcela do Corvo foi separada para esse fim, justamente os textos anotados em desvio. Utilizamos nosso aplicativo, com a tipologia de desvios embutida, e as ocorrências capturadas nos dão o seguinte panorama:

Desvios Gramaticais VP	53	160 Textos
Desvios Gramaticais FN	37	

Por ordem de frequência, as categorias e os tipos específicos de desvios verdadeiros (VP) formam o quadro abaixo:

Tabela V-6: Quadro geral das detecções dos desvios gramaticais

Categoria de Desvio	Tipo de Desvio	Etiqueta	Ocorrências
Concordância	concordância verbal	cov	18
Concordância	concordância nominal	com	8
Pontuação	uso de pontuação	ptn	7
Crase	uso de crase	cra	5
Funções e uso	uso de artigos e determinantes	det	4
Colocação pronominal	colocação pronominal	cop	2
Concordância	concordância modo e tempo verbais	cmt	2
Regência	regência verbal	reg	1
Funções e uso	uso de conjunções	cjc	1
Funções e uso	uso de mau/mal	mal	1
Funções e uso	uso de onde/aonde	ond	1
Funções e uso	uso de por que	prq	1
Verbos	uso dos verbos	ver	1
Verbos	uso de particípio	ptp	1
TOTAL			53

Consideradas pela perspectiva da frequência, as cinco primeiras categorias de desvio revelam quais são os assuntos gramaticais de mais difícil assimilação pelos nossos redatores – concordância, pontuação, crase e funções e usos (de elementos sintáticos). Pela perspectiva da variedade, as detecções mostram que esses redatores muito raramente incorrem em desvios do tipo pontual (p.ex., uso de mau/mal, uso de particípio).

É preciso ter claro que os textos do Corvo não são exemplos de ampla exposição de idéias, de rica argumentação, com intertextos e vasto repertório cultural. Não há riqueza vocabular, portanto, nem construções sintáticas sofisticadas, de modo que no quadro de detecções acima é compreensível a concentração dos desvios em categorias tidas como básicas na norma padrão. Falha-se mais na concordância do que na regência, porque o uso de expressões duvidosas quanto à regência é raro nos textos. Para ser possível produzir um desvio de concordância verbal, por exemplo, basta empregar um verbo na frase; o mesmo não se pode dizer de certos casos de regência. Vejamos alguns exemplos de ocorrências no contexto em que os desvios foram produzidos:

Tabela V-7: Quadro geral das detecções dos desvios gramaticais

Tipo de Desvio	Ocorrência de Desvios VP
Uso de particípio Concordância verbal	So existe oportunidade no mercado de trabalho para aqueles que estão <vp_ptp>melhor preparados</vp_ptp>... Pois nós queremos apenas uma oportunidade de igualdade e mais justo. Para com que <vp_cov>todos possa viver</vp_cov> e lutar pela uma vida melhor mais digna. [C2C_17]
Concordância verbal Uso de crase	Sabendo-se que <vp_cov>todos tem</vp_cov> essa proposta, com o decorrer do tempo vemos que é uma farça, (...) o homem se queixa do seu proprio ato <vp_cra>as vezes</vp_cra> não sabe nem pensa antes de agir e ficar a murmurar. [C2C_51]
Uso de pontuação	O povo brasileiro <vp_ptn>, parece</vp_ptn> que não ama mais seu país, vota em qualquer um, e é isso que dá, o Brasil quase falecido. [C3C_17]
Colocação pronominal	Más será se entrar um politico honesto que pensa em ajudar a população as pessoas lá dentro <vp_cop>deixa ele</vp_cop> ficar lá, eu acho meio difício. [C3C_22]
Concordância modo e tempo verbais	Só assim os mesmos <vp_cmt>podem transformarem</vp_cmt> o Brasil em um país de dignidade e justiça, para que também possam atender a sociedade em geral. [C2C_40]
Uso de pontuação Uso de crase Concordância nominal	O nosso país <vp_ptn>, é</vp_ptn> um país democrático. O direito de votar hoje, é concedido <vp_cra>à todos</vp_cra> nós cidadãos. <vp_com>O jovens</vp_com> apartir de seus 16 anos, já tem direito ao voto. [C3C_5]

A concordância do verbo “ter” (no plural, “têm”), como aparece nesse exemplário, foi negligenciada quase sempre; muitas vezes não foram detectadas, selecionadas para a anotação, mas é um fato que, nos textos do Corvo, poucos são os casos em que se aplicou a concordância como rege a norma padrão. Ora, em textos onde a acentuação gráfica elementar é fortemente desprezada, o que dizer dos acentos diferenciais?

Aliás, esse assunto da acentuação nos remete à crase. As detecções revelaram que a aplicação indevida da crase é mais freqüente do que a não aplicação dela. Curiosamente, enquanto nossos redatores parecem desprezar os acentos gráficos comuns, com a crase acontece justamente o contrário: há um uso indiscriminado desse acento. Em apenas um caso, e de falso negativo, houve confusão entre o “a” craseado e “há”. O que parece haver, de fato, é o desconhecimento das regras de emprego da crase que recomenda, por exemplo, o não uso do acento antes de pronomes pessoais, de pronomes indefinidos e antes de números, tal como ocorreu no corpus.

Nessa linha de subespecificação dos tipos de desvios, devemos considerar o importante tópico da pontuação. Podemos dizer, sem medo de errar, que o emprego dos sinais de pontuação, em especial a vírgula, parece se dar de modo aleatório, tal é a ausência de critérios (precisos) que

se observa no corpus. A pontuação está entre os tipos de desvios verdadeiros mais frequentes e o mesmo acontece no levantamento dos falsos negativos. Em todo o corpus, aliás, as reticências foram aplicadas apenas três vezes e num único texto; o sinal de exclamação, quando corretos, aparecem nos títulos. Interessante assinalar, com respeito à exclamação, que em alguns textos, especialmente em trechos que se quer ênfase na idéia, a exclamação aparece duplicada ou triplicada. Quanto ao ponto de interrogação, a grande questão parece ser: ele ainda é válido? No corpus, o que vimos é a ausência quase total do ponto em frases interrogativas. Observe-se, no seguinte exemplo, o que se verifica com frequência nos textos:

E o povo brasileiro o que faz! Nada! pois pensam, somente no seu próprio bem estar, bens materiais, etc. (C3A_6)

Mas seguramente a vírgula é com a qual se tem mais dificuldade. Um dos principais recursos para imprimir ritmo ao texto, as regras básicas da aplicação da vírgula também parecem ser desconhecidas pelos nossos redatores. Nas detecções de desvios verdadeiros, a maior frequência é a de uso indevido da vírgula, porque aplicada entre sujeito e verbo (4 casos). Os demais acusam a não aplicação da vírgula antes de conjunções em frases subordinadas ou explicativas (antes de “mas”, “pois” e “ou seja”, p.ex.).

Vejamos, agora, a tabela de ocorrências dos falsos negativos e, em seguida, alguns exemplos de ocorrências.

Tabela V-8: Quadro geral das detecções dos desvios gramaticais

Categoria de Desvio	Tipo de Desvio	Etiqueta	Ocorrências
Concordância	concordância nominal	com	15
Concordância	concordância verbal	cov	10
Pontuação	uso de pontuação	ptn	9
Regência	regência verbal	reg	2
Crase	uso de crase	cra	1
TOTAL			37

Tabela V-9: Quadro geral das detecções dos desvios gramaticais

Tipo de Desvio	Ocorrência de Desvios FN
Uso de pontuação	Vamos todos votar, devemos escolher os nossos representantes políticos: <fn_ptn>Governador, Senadores, Deputados Estaduais e Deputados Federais</fn_ptn> e Presidente (...). [C2C_44]
Concordância nominal	Por have essa desigualdade <fn_com>social vários</fn_com> conflitos gera em torno disso (...). [C2C_32]
Uso de pontuação Concordância nominal	(...) com essa onda de criminalidade é com esse aumento de corrupção, que cada vez <fn_ptn>mas</fn_ptn> empobrece nosso país. Temos que lutar pelos nossos direitos de cidadão fazer valer o nosso voto, porque tem casos de pessoas analfabetas, trocando voto por <fn_com>comida e várias outras coisas</fn_com>. [C3A_6]
Regência verbal	(...) então é por isso que todos nós devemos <fn_reg>votar para</fn_reg> que o nosso Brasil se torne um país digno (...). [C3B_9]
Concordância verbal	Quando nós transformamos em jovens temos o direito de escolhermos alguém para administrar o nosso país, mas <fn_cov>muita das vezes não queremos</fn_cov>, pois nós temos mais esperanças de que um dia alguém virar fazer algo para melhorar. <fn_cov>Tem jovens</fn_cov> que não se enteresa pela politica (...). [C2C_31]

O quadro de falsos negativos, de certo modo, mantém as categorias de desvio apontadas no levantamento dos verdadeiros positivos. Houve inversão das duas primeiras categorias, com a concordância nominal aparecendo, aqui, na primeira posição. Mas isso se justifica facilmente. É que, como já dissemos em outro momento, a palavra “país” não foi acentuada muitas vezes; com isso, uma construção freqüente no corpus, que é “nosso país”, foi tomado como um desvio de concordância pelo revisor. A falha na ortografia também é responsável por outros tipos de falsos negativos. Veja-se, no exemplário acima, a ocorrência de “mas”, que deveria ser “mais”, levando o revisor a interpretar equivocadamente o caso como um desvio de pontuação. Por sua vez, o uso precário da pontuação, tal como discutimos atrás, também conduz o revisor a detecções como a que vemos na ocorrência “social vários” – um falso negativo de concordância nominal. A título de teste, aplicamos a vírgula após a palavra “social” e, como prevíamos, o revisor não mais considerou o segmento um caso de desvio.

A falta de preparo do revisor para tratar adequadamente alguns fenômenos da língua (p.ex.: as ocorrências de concordância nominal em C3A_6 e de concordância verbal em C2C_31) é um outro fator de detecção indevida pelo revisor. Mas o que devemos ressaltar é a presença de regras de correção no revisor que parecem ser pouco flexíveis e não abrangentes. Foi o que se verificou na análise das ocorrências ligadas à regência verbal. Nas duas detecções de falsos negativos, o caso foi o mesmo: “votar para”. O revisor sugere “votar em”, ignorando o

complemento possível e correto do verbo introduzido pela preposição “para” (ex.: “Ele quer *votar para* garantir um futuro melhor”).

A propósito desse viés de discussão, vamos trazer ao debate os falsos positivos ou omissões. Como nosso leitor sabe, por decisão de projeto não realizamos a detecção exaustiva das omissões e não temos, portanto, o levantamento de dados correspondentes. Os textos são de difícil revisão, com presença de diversos problemas no plano de expressão, mas fora do escopo da nossa pesquisa.

Apesar disso, como reportamos na seção metodológica, foi considerado importante selecionar algumas ocorrências de falsos positivos, a fim de que pudéssemos interferir na tipologia de desvios com a qual trabalhamos no corpus. Interrogamo-nos, desde o início: essa tipologia dá conta do corpus? Ela consegue mapear os desvios gramaticais desses textos? Em que assuntos gramaticais ela poderia ser expandida, refletindo o corpus Corvo que lhe serve de base?

A resposta que damos, agora, é: o corpus é muito mais complexo do que supõe essa tipologia. Há fenômenos lingüísticos tão limítrofes entre o plano lexical-discursivo e o plano gramatical que uma rotulação assaz padronizada não pode ser fácil e, talvez, seja mesmo uma tarefa improdutiva.

A tipologia de desvios que trouxemos para tratar o Corvo só não foi plenamente satisfatória, porque é baseada na tipologia de erros de um revisor automático que provou, aqui, ser limitado. Mas, caso fosse maior a capacidade de processamento desse corpus pelo sistema, certamente veríamos importantes assuntos da norma padrão representados na tipologia.

Ressalte-se, ao mesmo tempo, que essa mesma tipologia é válida sim, desde que contém tópicos gramaticais que puderam comprovar-se freqüentes no Corvo, tais como a concordância e a pontuação. Em outras palavras, seria possível tornar os textos do corpus mais adequados à norma padrão e, conseqüentemente, mais apropriados à variante culta do português, se interferíssemos no texto justamente nos casos de desvios de concordância, pontuação, crase, etc., como se propõe na tipologia. Dito ainda de outro modo, a tipologia de desvios que utilizamos é válida, mas incompleta; é restrita, mas suficiente para tratar os desvios gramaticais presentes em textos como os do perfil do nosso corpus.

A título de contribuição, nossa análise sobre o corpus sugere alguns casos que poderiam ter sido contemplados no funcionamento do ReGra; outros que participam das categorias de

desvios já tratadas pelo revisor, mas são tipos de desvios específicos que nós, como usuários da ferramenta, supomos ser importantes se se quiser equipar o revisor para lidar de forma mais eficaz com textos como os do Corvo. Casos, enfim, que em vez de nos dirigirmos ao revisor, remetemos às fontes de formulação da norma padrão, que se tornam referências sobre a matéria no país, a fim de que reconsiderem o rigor da norma sobre o que efetivamente se produz na escrita contemporânea; dito de outro modo, que levem em conta uma análise baseada em corpus para decidir o que seria ou não correto na variante culta do português, ainda que, no âmbito deste estudo, possamos falar de uma pequena parcela dos usuários (aprendizes) do português culto.

Vejamos essas ocorrências de omissão, mostrando, primeiramente, o quadro de distribuição pelas categorias de desvio:

Tabela V-10: Quadro geral das detecções dos desvios gramaticais

Categoria de Desvio	Tipo de Desvio	Etiqueta	Ocorrências
Sem especificação na tipologia	-----	ses	8
Pontuação	uso de pontuação	ptn	5
Regência	regência verbal	reg	4
Concordância	concordância verbal	cov	3
Crase	uso de crase	cra	1
Funções e uso	uso de preposições	pre	1
Funções e uso	uso de pronomes	pro	1
TOTAL			23

Os casos sem especificação são justamente as nossas sugestões de revisão da tipologia. Os demais, como se vê, estão contemplados na tipologia, mas representam ocorrências que, segundo as mesmas referências de norma padrão utilizadas para o seu desenvolvimento, poderiam ser consideradas desvios gramaticais; em síntese, o revisor se omitiu diante de ocorrências que, sob o olhar de um especialista da norma padrão, seriam desvios gramaticais verdadeiros positivos. A relação completa desses casos da tabela acima está no Anexo II, já anunciado neste trabalho. Estão lá, porque além de sugerirmos a detecção e a classificação gramatical, também nos propusemos à formulação do texto explicativo que deve acompanhar a correção do desvio pelo revisor. Interessante é notar que, caso fossem acatados pela ferramenta, teríamos um regra de correção baseada em ocorrência extraída de um corpus em vez de baseada nos manuais de gramática.

De todo modo, vamos trazer algumas ocorrências, especialmente de etiqueta “ses”, e satisfazer o nosso leitor com mais esclarecimentos sobre essa tal contribuição que julgamos estarmos oferecendo com o nosso estudo. As ocorrências são mostradas ao lado de uma sugestão para intitulação do assunto gramatical (o tipo de desvio) e para a frase de correção.

Tabela V-11: detecções dos desvios gramaticais FP (ses)

Ocorrência	Tipo de Desvio (etiqueta)	Recomendação
O Brasil necessita de uma pessoa onesta que tenha dignidade <fp_ses>e sinsseridade e lealdade</fp_ses> com os eleitores, porque (...). [C2A_2]	Uso de conectivos (ses)	Verifique o excesso ou a falta de variação de conectivos.
(...) seria a falta de dignidade de muitos, <fp_ses>porque enquanto muitos estão com a barriga cheia, outros estão passando fome; devido ao fato de politicagem do Brasil, porque uns</fp_ses> ganham muito e tantos outros estão ai (...). [C3B_2]		Evite uma seqüência de conectivos de igual função na frase. [A palavra “porque” aparece mais de uma vez neste trecho]
Mudar a realidade que acabamos de <fp_ses>se-acustumar</fp_ses>, durante anos em que não podíamos fazer nada (...) virão e precisarão de um pouco de paz e condições de <fp_ses>se-adaptarem</fp_ses> nesse lugar (...). [C2B_7]	Emprego do hífen (ses)	O hífen não se aplica neste caso
Hoje o voto é muito importante para <fp_ses>tentar-mos</fp_ses> mudar o destino de um país como o Brasil pois este direito (...). [C2C_59]		O hífen não se aplica neste caso
Seu direito de votar é essencial, com ele mudamos uma nação inteira, <fp_ses>então conclui-se</fp_ses> que o voto é a maior arma (...). [C3B_5]	Redundância (ses)	Evite o uso de expressões redundantes em cadeia
(...) colocando no poder pessoas incompetentes e sem caráter que <fp_ses>irão levar</fp_ses> o nosso amado Brasil ao fundo do posso, mais do que já está. [C1C_6]	Tempo verbal (ses)	Elimine o verbo auxiliar, empregando o verbo principal no futuro simples
Para com que todos possa viver e lutar <fp_ses>pela uma</fp_ses> vida melhor mais digna. [C2C_17]	Contração de preposição (ses)	A contração da preposição não se aplica neste caso
Eu creio que, na época da politica é o momento em que <fp_pro>nós se tornamos</fp_pro> muito especial para os políticos. [C2C_42]	Uso de pronomes (pro)	Empregue o pronome relacionado ao verbo
(...) mas também quando votar, não podemos pensar só em nós. <fp_reg>Temos que</fp_reg> pensar em todos.	Regência verbal (reg)	Verifique o tipo de complemento exigido pelo verbo
(...) todos nós devemos votar para que o nosso Brasil se torne um país digno sem desemprego, <fp_ptn>sem fome sem violência</fp_ptn>, sem meninos de rua, um país sem guerra, etc. [C3B_9]	Uso de pontuação (ptn)	Empregue a vírgula para separar os integrantes da enumeração

Tabela V-11: detecções dos desvios gramaticais FP (ses) (cont.)

Ocorrência	Tipo de Desvio (etiqueta)	Recomendação
Temos <fp_ptn>que, ampliar</fp_ptn> nossa mente, parar de pensar o que aquele governante vai fazer pelo meu estado (...). [C3C_20]	Uso de pontuação (ptn)	Neste caso a vírgula não deve ser empregada
<fp_cov>Metade da sociedade brasileira vendem</fp_cov> seu voto, como eleitor bom e verdadeiro o povo poderiam pensar que isso não é ser honesto. [C2C_40]	Concordância verbal (cov)	Verifique a concordância verbal. [Se "vendem" estiver se referindo a "metade da", verifique a concordância]

O que temos a dizer, para encerrar, é que, não obstante a falta do levantamento preciso da frequência de ocorrências, esses tipos de desvios acima listados são bastante recorrentes no corpus. Mais que revelar qual desses é o tipo de desvio mais frequente é termos claro que são nesses casos que um redator como o nosso encontra maior dificuldade. É compreender, por outra parte, que, uma vez recorrente, desvios como o da concordância e regência verbal apontados nesse quadro são já bastante comuns na escrita contemporânea (para não dizer na fala), de modo que, na contramão da gramática tradicional, se pode admitir a ocorrência menos como um desvio, mas como uma tendência para a mudança lingüística que se consolidará numa nova era do português culto.

V. 3 A superfície textual e o processo de produção da escrita

Trabalhar com textos originais manuscritos nos impôs severas dificuldades e algumas frustrações, como a perda de arquivos, o lento processo de transposição de registros, etc. Mas, como nosso leitor deve ter notado, cedo decidimos transformar aquilo que seria uma simples tarefa de compilação em algo mais denso, revelador de informações pouco comuns num trabalho com corpus. Acreditamos que essa decisão trouxe vantagens para o nosso estudo, como veremos agora.

Falamos da análise daquilo que chamamos de *superfície textual*, i.e., a apresentação do texto, seu *layout*, o aspecto lingüístico que está por trás da escrita propriamente dita. Como se sabe, elegemos alguns traços característicos dos nossos textos para realizarmos um levantamento preciso, submetendo-o à análise depois.

Esses traços estão separados em oito categorias, e a descrição de cada uma, bem como o processo de inserção das etiquetas nos textos, puderam ser conhecidas no capítulo anterior. O que trazemos, agora, é o resultado do mapeamento dessas etiquetas de superfície.

Iniciamos com a apresentação do quadro geral de frequência, conforme a tabela abaixo:

Tabela V-12: Frequência dos traços da superfície textual observados no Corvo

Superfície Textual	Ocorrências (A)	% Ocorrências (sob total em A)
Qualificação do título	41	16
Hesitação	39	16
Inserção	19	8
Apagamento	13	5
Ilegibilidade pontual	13	5
Assinatura	3	1
Ruído	2	1
Cabeçalho	1	0,4
TOTAL	159	

E de imediato, lancemos ao leitor outro dado: a distribuição dessas ocorrências no corpus. Por pertinente, apresentamos os dados já apontando a medida da concentração das ocorrências:

Tabela V-13: Distribuição da presença dos traços da superfície textual no Corvo

Ocorrências	Textos	% Textos
Nenhuma	125	50
Uma	96	39
Duas	23	9
Três	3	1
Quatro	2	1
TOTAL	249	100

O que se observa, de partida, é que metade do corpus não apresentou especificidades do tipo que investigamos, mas que, na outra parcela do corpus, são cerca de 160 ocorrências pertinentes, grande parte delas com potencial de nos revelar importantes informações a respeito do processo de escrita.

Começamos pelo aspecto do título. A anotação dessa especificidade visou ao levantamento do quão freqüente é a reprodução do tema proposto ao redator que, ao fazê-lo, deixa de cumprir uma expectativa de seu leitor e de realizar essa etapa tão importante que é

intitular seu texto. Vemos, então, que 16% dos redatores simplesmente repetem o tema proposto no ENEM, levando o seu leitor, desde o título, a uma avaliação se não precária, ao menos desafortunada do texto.

A frequência dos cabeçalhos e assinaturas no corpus é baixa, o que mostra que quase a totalidade dos nossos redatores está atenta à estrutura padrão do tipo de texto exigido no ENEM. De fato, em textos para avaliação escolar não cabe identificação de qualquer tipo.

Considerados esses aspectos, passemos àqueles que implicam maior reflexão. Conforme discutimos na seção teórica, alguns setores do sistema escolar vêm insistindo na questão da boa apresentação do texto, especialmente se o redator está sob avaliação. Recomenda-se, em linhas gerais, uma caligrafia razoável, a estruturação do texto em parágrafos tabulados, a ausência de rasuras, enfim, deseja-se aquilo que se intitulou “texto limpo”. Isso porque o texto tem um leitor criterioso, empenhado em capturar deslizos de qualquer espécie e, desse modo, um texto insatisfatório na superfície pode prejudicar o desempenho do redator na avaliação. Foi justamente para medir o grau de adequação dos nossos redatores nesse quesito, que detectamos as ocorrências de ilegibilidade e ruído.

Dos redatores que estão nesse universo dos que algum problema acusa em termos de superfície textual 5% pioram sua situação apresentando ilegibilidade pontual. É um número expressivo, uma vez que falamos de um trecho do texto que ofereceu grande dificuldade ao leitor, e talvez tenha desviado a sua atenção, que poderia estar concentrada em saber a proposta do texto, saber o que pensa o redator.

A ampla maioria das redações com as quais trabalhamos realmente são de difícil leitura, no sentido de que exigem grande esforço em decifrar o que está escrito. É forçoso dizer que se buscou o máximo de parcimônia possível com os textos durante a transposição de registro, sendo considerada uma ocorrência de ilegibilidade, por exemplo, aquilo que verdadeiramente não foi desvendado com a leitura. Fomos benevolentes com o redator, em certa medida, pois algumas vezes, em casos de poucas dúvidas, acabávamos agindo em favor do acerto do redator e não o contrário (p.ex.: digitar uma inicial maiúscula em sentença, mesmo parecendo ser a minúscula a que estava no manuscrito). Quando dizemos, então, que 13 ocorrências é um número significativo, o que enfocamos é o descaso desses indivíduos com o seu leitor, pois, não bastasse o fato de talvez produzirem um texto cujo conteúdo não correspondesse à tarefa proposta, não

bastasse ainda um leitor disposto a decifrar uma caligrafia ruim, mesmo com isso ele não se preocupou com a construção de um texto limpo. Quanto ao traço do ruído, que está nessa linha do prejuízo da leitura, as duas ocorrências podem ser consideradas um incidente, até mesmo de digitalização do manuscrito; de todo modo, ele não foi recorrente, sendo inexpressivo para a análise.

Consideremos, agora, três especificidades dos nossos textos as quais, na contramão do que indicavam as outras, são fatos que podem levar o leitor a valorizar o texto. São as marcas de inserção, de apagamento e de hesitação.

Dissemos que uma ocorrência de inserção se dá, quando uma palavra ou uma sequência de palavras é enxertada em determinado trecho, consistindo numa ação posterior ao fluxo inicial da escrita. O apagamento é a eliminação de palavra ou parte dela ou mesmo uma sequência de palavras com intenção de não ser considerada na leitura. A hesitação, por fim, é o caso em que o redator interfere pontualmente naquilo que escreveu, promovendo alterações que influenciarão no seu desempenho quanto à norma padrão. Como se vê, essas especificidades se relacionam ao processo de produção escrita, à formulação do texto. E capturar essas propriedades representa tentativa de mostrarmos indícios de que os nossos redatores, quando apresentam essas marcas, revelam-se preocupados com o desvio, receosos em não manter o plano inicial de escrita, buscando rever determinadas construções, palavras ou sinais gráficos lançados ao texto num primeiro momento.

O resultado do levantamento desses três casos, tal como está na primeira tabela, foi bastante surpreendente. Das ocorrências de superfície, 30% concentram-se nessas categorias; mais da metade desse número se refere à hesitação. O que esses resultados nos revelam, de princípio, é que há seguramente uma tendência dos nossos redatores de revisar a escrita, reconhecendo, mais que qualquer outra questão, a existência de um padrão na variante culta. Tanto mais pelo fato de sabermos que nesses tipos de ocorrências a alteração que promovem são em favor do acerto, ou seja, apaga-se, insere-se ou hesita-se acertando justamente na aplicação da norma padrão ou na ortografia oficial. Vejamos alguns exemplos de ocorrências:

Tabela V-14: Exemplos de ocorrências no mapeamento da superfície textual do Corvo

Marcas da Superfície Textual	Ocorrências
Inserção	(...) quanto mais o tempo passa, não dá para as pessoas verem <in_lexgr>que</in_lexgr> com as experiências (...). [C3C_25]
	(...) povo que vive as escondidas sem saber como <in_gra>se</in_gr> pronuncias (...). [C3C_31]
	(...) somos obrigado a votar, e escolher pessoas incapazes de nos representar na câmara, prefeitura, assembleia e <in_lex>no</in_lex> senado. [C5C_6]
	Mas uma obrigação que agente <in_lex>tem</in_lex> com o nosso País (...). [C3C_96]
Apagamento	(...) mas a reflexão é um <ap_sbr_lex>bom</ap_sbr_lex> método (...). [C3C_12]
	Dessa forma as pessoas não <ap_sbr_cadeia>pensam e</ap_sbr_cadeia> análise os candidatos (...). [C3C_31]
	(...) de uma democracia onde <ap_sbr_lex>você</ap_sbr_lex> a pessoa é a única responsável (...). [C3C_94]
Hesitação	(...) para contruirmos uma <h_ort>nasão</h_ort> nação completa com (...). [C2C_48]
	(...) o voto é obrigatório e com <h_num>eles</h_num> ele vamos (...). [C3C_15]
	A escolha é muito importante, pois é <h_ac>está</h_ac> esta que ditará as regras da eleição (...). [C3C_29]
	(...) esse direito, que foi conquistado <h_ort>atravéz</h_ort> através de uma história homérica (...). [C3C_85]
	(...) pois se o país tem a chance de mudar, basta <h_hf>aproveitála</h_hf> aproveitá-la muito bem. [C4C_25]
Conjuntas: hesitação e inserção	(...) sabemos também a existência de dois tipos de votos o <h_ort>conciente</h_ort> consciente e o inconsciente. Quantas pessoas, <in_cadeia>no passado,</in_cadeia> lutaram por este direito e hoje (...). [C3C_55]
Conjuntas: apagamento e inserção	(...) mas <ap_sbr_lexgr>só</ap_sbr_lexgr> pessoas como o Srº Collor de mello, acabam <in_lex>decepcionando</in_lex> (...). [C4C_16]

Note-se que propositadamente as ocorrências de apagamento são do tipo “sombra”, i.e., as que de algum modo foi possível descobrir o que se pretendia esconder. Atente-se, além disso, para as ocorrências conjuntas, i.e., que aconteceram num mesmo texto. Elas mostram que esse redator buscou a revisão por mais de uma via, numa postura que certamente merece ser levada em conta.

Ocorreu-nos, por pertinente, que esse grupo de textos com presença de marcas de apagamento, inserção e hesitação fossem relacionados ao mapa dos textos na questão dos desvios. A questão que formulamos foi: em que medida os redatores que incorreram em desvios apresentaram textos falhos na superfície?

Primeiramente é importante ter claro que enfocamos o grupo de textos com co-ocorrência de etiquetas de superfície e buscamos relacioná-lo ao grupo de textos também com co-ocorrência,

mas a de desvios dos tipos ortográficos e gramaticais, ambos verdadeiros positivos. Temos, de início, um cenário que se mostra da seguinte maneira:

Figura V-3: Quadro de distribuição de desvios e marcas de superfície

Esses resultados nos apontam: a) que os textos com desvios combinados – ortografia e gramática – representa quase a totalidade do corpus: 98% dos textos; b) que se há desvio gramatical num texto, a chance de esse texto também apresentar um desvio ortográfico ou vice-versa é de mais de 50%; c) que mesmo não co-ocorrendo, os desvios estão presentes em quase a metade do corpus; d) que dos textos com marcas de inserção, apagamento ou hesitação, em cerca de 1/3 deles há ocorrências combinadas e outra quantidade dessas com ocorrências de apenas um dos fatos observados.

Prosseguimos nessa linha de investigação e empreendemos uma análise da distribuição conjunta desses dois grupos de textos com co-ocorrência. Desta vez, o gráfico obtido apresentou-se da seguinte forma:

Figura V-4: Distribuição conjunta de desvios e marcas de superfície (co-ocorrência)

Esses resultados indicam de pronto: a) que em relação ao grupo dos textos com desvios não co-ocorrentes, o grupo de textos com desvios co-ocorrentes apresentou maior concentração de ausência de marcas da superfície – 55% versus 44%, respectivamente; e b) que o grupo de

textos com desvios não-recorrentes apresentou uma incidência 10 pontos percentuais maior de presença de marcas de superfície textual (desde que sejam marcas não-recorrentes) do que o grupo de textos com desvios recorrentes – 30% versus 20%, respectivamente.

Significa dizer, a partir desse quadro, que os textos em que aparecem desvios combinados é pior a qualidade da superfície textual, já que nenhuma ocorrência das marcas de superfície que poderiam beneficiar o redator (as do gráfico em questão) ocorre nesses textos. Quando o texto apresenta desvios individuais (ou ortográficos ou gramaticais), sabe-se que o redator empenhou-se numa revisão da escrita (56% dos indivíduos ou hesitaram em se desviar ou apagaram trechos ou inseriram outros), estando em maior número aqueles que realizaram essa revisão deixando marcas de estratégias não combinadas (ou só via apagamento ou só inserção ou só hesitação).

Da perspectiva de quem se preocupa em saber como é o processo de assimilação da norma padrão pelos redatores como os do nosso perfil, importa destacar com esses dados que parece haver indícios de que o *aprendiz* mais bem preparado na norma padrão é também mais consciente do que é o processo de escrita, sabendo que pode realizar, e de fato realiza, uma revisão do texto. Talvez seja um caminho, então, não só aparelhar esse aprendiz com as regras prescritivas, mas também construir nesse indivíduo uma percepção de que, ao menos em textos desse gênero, a simples ação de voltar ao que antes estava escrito e alterar se for preciso contribui sobremaneira para a adequação do texto à norma padrão e, quiçá, com um desempenho mais positivo num contexto de avaliação.

E quem será esse aprendiz? Tais afirmações baseiam-se em quais indivíduos? Suspendemos a exposição neste momento para abordar o perfil de indivíduos para os quais falam esses dados do nosso corpus.

V. 4 Perfil sócio-econômico dos redatores: interpretações na contramão do ENEM

Nesta pesquisa, associar os traços lingüísticos sobressaltados nos textos com a caracterização sócio-econômica dos redatores é algo que esteve no horizonte desde o princípio. Tivemos a oportunidade de adquirir, junto com os textos, informações valiosas sobre os autores das redações do ENEM. Pudemos conhecer sobre eles, por exemplo, em que faixa de renda se situam, o que poderia nos indicar se se tratava de um indivíduo cuja situação financeira lhe permitisse acesso a recursos importantes, garantindo-lhes um ambiente cultural que, em grande

medida, interfere na sua formação escolar, na construção do repertório linguístico do tipo que está em foco aqui.

É bem verdade que, ao nos vermos diante de um banco de dados tão rico, gigantesco e que tão poucos têm acesso, nos sentimos frustrados por estarmos aquém do perfil de pesquisa/pesquisador que esse material merece ter. Nosso corpus não deixou de ser expressivo; com todas as baixas na seleção e compilação textual, reunimos cerca de 250 textos os quais, aliás, têm uma diversidade de perfil proporcional à população do ENEM. No entanto, não se pode deixar de admitir que a base de dados que temos como referência imediata (a Base-Corvo) conta com aproximadamente 700 mil textos; a que vem em seguida (a Base-ENEM) têm a incrível soma de 1 milhão e 200 mil unidades aproximadamente. Sem sombra de dúvidas, portanto, essa nossa fonte de pesquisa ainda ficará para ser bastante investigada.

Nossa meta de trabalho, a esse respeito, foi muito bem delimitada desde o início: estudarmos os textos mais mal avaliados no ENEM, e somente eles. E sabíamos, pela experiência profissional no âmbito do próprio ENEM, que essa população não seria pequena, de modo que, em certa medida, estaríamos nos ocupando de uma massa de textos/indivíduos que é expressiva no Exame (a Base-Corvo representa 46% do ENEM). Então, incidiu sobre as decisões de projeto o recorte específico da nossa investigação, que foi, dentre os mais mal avaliados, focalizar aqueles textos nas faixas inferiores da avaliação no quesito “domínio da norma culta” (para nós, adequação à norma padrão).

Nossa base com 249 reflete, em termos de *perfil* de redator e de nota obtida no Exame, a população de *baixo rendimento* nesse quesito de avaliação, seguindo uma escala que, além dessa, possui a população de rendimento médio (ou expressivo) e a de rendimento superior (ou superação). Nesta última faixa estão os casos em que o texto recebeu a nota máxima na avaliação da adequação à norma. Mas, tal como pudemos comprovar durante a investigação, dentre os quesitos avaliados no Exame, a adequação à norma padrão é de longe o mais positivo, mostrando que, de certo modo, os avaliadores agiram com peso diferenciado na atribuição de notas nesse quesito. Certamente teríamos uma massa populacional muito maior de referência para ao corpus, se nos propuséssemos a investigar os casos de superação, por exemplo.

A bem da verdade, as notas atribuídas aos textos não mais nos inspiraram tanta confiança, a partir dessa constatação. Daí, por exemplo, não nos concentrarmos na compilação apenas dos

textos auferidos como de rendimento insuficiente (Ri), incluindo nesse plano de observação também os textos de rendimento baixo (Rb) e de rendimento médio (Rm). E agindo desse modo, pudemos garantir para essa base de 249 textos um perfil abrangente de desempenho na avaliação, já que reflete três (e não uma) faixas de rendimento no quesito norma culta do ENEM.

No perfil do corpus Corvo, a presença maciça de textos Rm é um reflexo dessa população de referência, que é a faixa de maior população na Base-Corvo. Mas isso também implicou sabermos, desde cedo, que menor seria o número de textos/indivíduos que, segundo o ENEM, realmente são de pior desempenho nesse quesito da norma. De certo modo, nossa avaliação dos textos competiu com a do ENEM para a qual os textos Rm foram considerados satisfatórios no domínio da norma culta. Para nós, como vimos, há uma tal caracterização textual desfavorável no corpus e de forma homogênea, i.e., trata-se também de uma propriedade dos textos Rm. E como veremos mais adiante, a incidência de desvios gramaticais verdadeiros em textos dessa faixa de rendimento é bastante significativa. Não pudemos perceber, assim, em que medida os textos Rm se diferiam tão explicitamente dos textos Ri e Rb, já que, para citar uma das conclusões já apontadas neste estudo, a competência para a ortografia é precária em quase todos os redatores (lembremos: cerca de 85% do corpus apresentou verdadeiros positivos de ortografia).

Não é nosso propósito direcionar toda a nossa pesquisa, a investigação empírica dos textos para o sistema ENEM, sua classificação dos textos e diferenciação dos redatores, posto que isso nos exigiria outro percurso e talvez outros materiais e referenciais teóricos. Contudo, neste ponto da nossa análise, não é possível deixar de confrontar os dados do ENEM e admitir, por exemplo, que os critérios que se adotam lá para dizer que um indivíduo apresenta ou não domínio da norma culta e em que medida são questionáveis.

De certo modo, uma tal visão desconfiada sobre a separação dos indivíduos pelo perfil de rendimento nos levou a observar quem são nossos redatores não mais por esse prisma do rendimento, mas pela perspectiva dos traços do perfil do redator. Isto é, decidimos, primeiro, observar o comportamento lingüístico dos textos relacionado os dados a cada um dos itens do perfil aqui enfocados: a idade, a origem, a renda e o tipo de instituição escolar.

Supondo que o levantamento dos dados gramaticais (e não os da ortografia, por exemplo) é o que poderia nos levar a identificar, talvez, um *padrão* de indivíduos com desempenho de fato ruim no uso da norma padrão, partimos dos dados da detecção dos verdadeiros positivos

gramaticais extraídos com a validação dos desvios pela via da plataforma Word, já que, como vimos, houve melhor desempenho do revisor na detecção dos desvios gramaticais via Word do que via IDv. Em outras palavras, temos um levantamento qualitativo do desempenho dos redatores no uso da norma padrão, e isso foi o ponto de partida para a nossa análise.

O quadro de distribuição dos verdadeiros positivos por cada um dos itens do perfil é o que se apresenta abaixo:

FiguraV-5 : Distribuição da presença dos traços da superfície textual no Corvo

Idade	Desvios Gramaticais VP*			
	Zero	Um	Mais de um	TOTAL
Até 18 anos	27	26	17	70
18 a 21 anos	53	46	38	137
Acima 21 anos	22	11	9	429
TOTAL	102	83	64	249

*valores em textos

Região	Desvios Gramaticais VP*			
	Zero	Um	Mais de um	TOTAL
Norte	7	2	3	12
Nordeste	39	19	12	70
Sudeste	44	46	30	120
Sul	9	8	13	30
Centro-Oeste	3	8	6	17
TOTAL	102	83	64	249

*valores em textos

Renda	Desvios Gramaticais VP*			
	Zero	Um	Mais de um	TOTAL
Nenhuma	3	0	5	8
Até 5sm**	84	64	43	191
Acima 5sm	15	19	16	50
TOTAL	102	83	64	249

*valores em textos **salário mínimo

Instituição	Desvios Gramaticais VP*			
	Zero	Um	Mais de um	TOTAL
Só pública	94	65	52	211
Maior pb**	2	3	5	10
Só particular	6	15	7	28
Maior pr***	0	0	0	0
TOTAL	102	83	64	249

*valores em textos

**Maior parte escola pública

***Maior parte escola particular

Na linha de análise que adotamos até aqui, cumpre observar nesses gráficos e tabelas que o revisor, tal como vem se confirmando, não funcionou de modo tão satisfatório para os textos do nosso corpus, já que em todos os itens do perfil é alta a incidência de nenhuma intervenção da ferramenta. Não devemos interpretar, portanto, que a ausência de ocorrências significa bom desempenho do redator no uso da norma padrão, mas, antes, que o revisor utilizado na detecção dos desvios não foi suficientemente capaz de tratar seus textos, seja porque há frequentes desvios ortográficos, seja porque muitos deles apresentam uma superfície textual precária, para nos atermos aos fatores de relevância discutidos até aqui.

No entanto, ainda que a ferramenta deixasse a desejar em termos de desempenho, o fato é que ela agiu com precisão num número alto de vezes, em alguns casos intervindo duas vezes ou mais num único texto. Levando isso em conta, podemos dizer, então, que no que diz respeito ao perfil dos redatores:

☞ a revisão automática *não* detectou corretamente desvios gramaticais (portanto, encontrou maior dificuldade) em textos de indivíduos: a) acima de 21 anos; b) provenientes da região Norte e Nordeste; c) com renda até 5sm; e d) provenientes da rede pública.

☞ o desempenho da ferramenta melhora nos textos dos indivíduos na faixa de idade intermediária – entre 18 e 21 anos –, em relação aos indivíduos concluintes do ensino médio à época do exame (i.e., até 18 anos em 2002), voltando a decair e mais severamente para os textos dos egressos do ensino médio (i.e., acima de 21 anos em 2002).

☞ embora a faixa de indivíduos provenientes do nordeste e sudeste seja maciça no corpus, é nos textos dos redatores do sul e centro-oeste que a ferramenta funciona melhor.

☞ o indivíduo que tem acesso ao ensino particular apresenta um texto para o qual a ferramenta é mais eficaz do que o que concentrou seus estudos na escola pública.

Quando entendemos que a eficácia do revisor implica maior detecção de desvios gramaticais verdadeiros, esses apontamentos acima adquirem novos contornos. Tal como discutimos no tópico sobre os desvios gramaticais, se o revisor funcionou melhor em determinados textos é porque a qualidade da ortografia e da estruturação sintática também era maior ali. Por um caminho, talvez, menos previsível de análise, o que se vê é que para auferir desempenho ruim aos nossos redatores e saber quem eles são foi preciso: a) esvaziar os dados oficiais do INEP com respeito às notas e classificação dos indivíduos em termos de rendimento; b) compreender que se trata de um conjunto homogêneo de indivíduos no sentido de que a ampla maioria deles produz um texto com falhas frequentes na ortografia e muitas vezes de impacto no plano gramatical, quiçá no discursivo; e c) ponderar a detecção gramatical, desde que ela foi promovida aqui por meio de um mecanismo automático, muito longe do que seria caso a detecção fosse realizada por um humano.

Foi preciso, sobretudo, admitir que não temos dados suficientemente confiáveis para generalizar o perfil de redatores a ponto de afirmar que, no país, é mais frequente, por exemplo, a presença de desvios gramaticais em textos de indivíduos de baixo poder aquisitivo ou indivíduos da região norte, e assim por diante. Ao final, o que apresentamos com o corpus é um forte *indício* de qual seria esse perfil típico de redatores, uma vez que, neste trabalho, esses redatores serão sempre uma parcela apenas dos aprendizes do português culto.

A nossa análise chega ao fim insistindo no fato de que não é válida a idéia de que os redatores escrevem *mal* o português; antes, eles escrevem do modo como foram capazes de compreender a norma padrão e a ortografia oficial do português e, como vimos, são sujeitos em grau crítico de conhecimento ou de assimilação dessas bases. Não temos redatores que *não sabem* o português, mas sim que não estão equipados e conscientes dos padrões da variante culta. Parafraseando Rosa Virgínia Mattos e Silva, o português não é um, mas são dois, três... quantos serão realmente?

Quanto ao ENEM, seu mecanismo de atribuição de notas e o impacto que tem sobre os seus candidatos, afinamo-nos com o discurso de Carlos Alberto Faraco: “Pensamos que tudo isso [o ENEM] acabará um dia nos tribunais. Como a referência geral do Exame é o padrão artificial

(e a falta de domínio deste implica em perda de pontos); e como já dispomos de gramáticas e dicionários que acolhem vários aspectos da norma culta real como padrão, os alunos prejudicados certamente irão, um dia, cobrar seus pontos na justiça! E aí, talvez, se abra uma discussão política conseqüente sobre a norma-padrão brasileira”.

VI. Considerações finais e perspectivas

Estudar o português, investigar as redações, observar os redatores, construir o corpus, desenvolver o aplicativo, enfim, este trabalho de pesquisa foi envolvente desde o princípio. Estabelecer o intercâmbio com informatas, estatísticos, membros de banca de avaliação, colegas professores e lingüistas, isso também constituiu uma experiência só possível num trabalho tão interdisciplinar como esse.

Desconfiávamos, e hoje vimos que tínhamos razão, de que trabalhar com um corpus de textos como essas redações nos ofereceria valiosa oportunidade de capturarmos algo que, como professor ou como avaliador de redações, apenas tínhamos como suspeita: esses textos não são descartáveis; essa escrita não é pobre; esses redatores não são fracassados.

Mas nosso sentimento é o de que poderíamos ter ido além. Mas para qual das intersecções de pesquisa? Qual dentre os tantos assuntos conectados neste trabalho deveria ser privilegiado? Não conseguimos distinguir. Qualquer temática de que tratamos aqui merece ser investigada a fundo, em detalhes, com precisão, riqueza de informações, confronto com outros dados, enfim, acreditamos que o que investigamos carece de muitas pesquisas, mais pesquisadores interessados nessas questões, sempre lembrando da natureza interdisciplinar do trabalho.

De todo modo, vamos apontar nesta seção final alguns pontos do pesquisa que poderiam ser aprimorados, algumas tarefas que mereceriam ser concluídas e algumas questões conceituais que deveriam ser mais enfatizadas.

Primeiramente, tomemos o corpus. A questão urgente que ele nos inspira é a da conclusão do seu *design*, ou seja, cumprir a última etapa que se recomenda num trabalho com corpus, que é a disponibilização. É uma tarefa complexa, e por isso fora do nosso alcance, uma vez que são necessários planejamento dos mecanismos de disponibilização, que exige estudo do acesso ao corpus – por quem, por onde, de que modo –, suporte para a implementação do sistema de disponibilização, implicando negociação com parceiros (inclusive institucionais) até para a questão do custeio. Mas antes de qualquer coisa nesse sentido, há de se enfrentar o desafio da

obtenção da licença de uso dos textos. Sem estar disposto a discutir a ética, mas também a contribuição científica da disponibilização desses textos (na íntegra, aliás), não haverá trabalho profícuo. E esse desafio, nos parece, é promissor na medida em que são muitos os tipos de pesquisa que podem se beneficiar do acesso ao Corvo.

Esse corpus, ao final, apresentou-se em duas versões, mas uma delas – a versão anotada – pode ser retomada a fim de que apareçam todos os desvios detectados, mas não etiquetados com o nosso trabalho. Isso sem mencionar a detecção das omissões, i.e., aqueles casos de desvios que não foram alçados pelo revisor, mas são detectados pelo humano. Os textos do corpus podem ter um levantamento muito mais exaustivo dos desvios (mas não só deles) do que o que promovemos aqui, se se propuser a essa finalidade especificamente.

Ainda quanto ao corpus, consideramos que o volume que ele tem hoje é significativo, mas em vista do banco de textos tão maior de onde provém, é de se levar em conta a idéia de expansão significativa do corpus, dobrando ou triplicando o seu tamanho. Muitos são os tipos de pesquisa que dependem de corpora de largas proporções, tais como as pesquisas em PLN, como destacamos neste trabalho, para a qual o corpus poderia funcionar como uma espécie de *benchmark*, um recurso para acesso comum.

Mencionar a expansão do corpus nos leva ao tema do perfil dos redatores e, a esse respeito, uma perspectiva de pesquisa que muito nos seduz é a dedicada a outros segmentos do perfil, outras faixas de rendimento ou nota. Mesmo uma investigação que permanecesse com o foco na competência 1, poderia confrontar os dados com a avaliação dos textos nos outros quesitos. Na nossa pesquisa, aliás, não abordamos qualquer um senão a o quesito da norma culta (norma padrão, na verdade).

A propósito da norma padrão, procuramos levantar questões que julgamos pertinentes, em voga nos dias de hoje com respeito ao uso de uma norma lingüística, quem usa, como usa, mas também como ela é formada e quem a transmite. Enfrentamos com dificuldades o desafio de estabilizar uma tipologia de desvio, justificando essa iniciativa com a pertinência que ela revela para determinados campos de estudo, como é o da revisão automática da escrita. Mas há de se discutir essa tipologia. Não investigamos o impacto não só dela, mas do próprio trabalho que realizamos com ela, sobre o sistema educacional, sobre o meio envolvido com o material didático, enfim, o assunto pode ser recolocado à luz de outros interesses de pesquisa.

Para finalizar esses breves apontamentos, devemos mencionar os nossos aplicativos. Ao tempo desta pesquisa, esses recursos – IDv e EDv – foram não só planejados, estudados, modelados, mas apenas parcialmente implementados. Há muito a ser feito neles/com eles, sobretudo para torná-los robustos, consistentes e confiáveis ao usuário. Usuário, aliás, que pode ser ampliado, do mesmo modo como podem ser expandidas as funcionalidades, reduzidas a dependência de língua e de tarefa, e assim por diante. Pensamos que outros revisores automáticos podem se beneficiar desses recursos, sendo eles e não mais o ReGra o processador dos textos. A pesquisa, nesse caso, levaria os aplicativos a tornarem-se também independentes de sistemas de revisão (na verdade, detecção e anotação de corpus). Há de se pensar, enfim, em tornar esses aplicativos de uso público, para que sejam reutilizados e também criticados.

Parece-nos, contudo, que o impulso inicial de todas essas perspectivas foi colocado neste trabalho. Ficaremos circunscritos a ele, por hora, na expectativa de que não por força das nossas observações, mas pelo bom traçado do nosso futuro, o português se note sem os preconceitos de hoje, o aprendiz da norma sem os tropeços tão condenáveis e excludentes, o ensino de língua materna realmente diferenciado, ocupando-se com a construção de um indivíduo de sucesso em qualquer que seja o contexto ou a situação de comunicação.

VII. Referências bibliográficas

- ARPPE, A. (2000) Developing a grammar checker for Swedish. *Proceedings of the 12th Nordic Conference in Computational Linguistics (Nodalida99)*. Department of Linguistics, Norwegian University of Science and Technology, Trondheim, p. 13-27.
- ATKINS, B.T.S., ZAMPOLLI, A. (1994) *Computational Approaches to the Lexicon*. New York: Oxford University Press.
- BAGNO, M. (2004) *Lingüística da norma*. São Paulo: Loyola.
- BERBER SARDINHA, T. (2000) Corpus Linguistics: history and problematization. *DELTA*, vol.16, no.2.
- _____ (2004) *Lingüística de corpus*. São Paulo: Manole.
- BIBER, D., CONRAD, S., REPPEN, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- BIDERMANN, M.T.C. (1978) *Teoria lingüística: Lingüística quantitativa e computacional*. 1^a ed. Rio de Janeiro: Livros Técnicos e Científicos, v. 1, 277 p.
- BONFANTE, A. G., NUNES, M. G. V. (2000) *Métodos empíricos para correção gramatical de línguas naturais*. Relatório técnico do ICMC-USP, 108, 63 p.
- BORBA, F.S., NEVES, M.H.M., IGNÁCIO, S.E. (2002) *Dicionário de usos do português do Brasil*. São Paulo: Ática, v. 1, 1.674 p.
- BRASIL (2002) Ministério da Educação e Cultura (MEC). *Relatório pedagógico do ENEM – 2002*. Disponível em www.inep.gov.br . Acesso e download em jan/2005.
- CASTILHO, A.T. (2004) Variação dialetal e ensino institucionalizado da língua portuguesa. In: BAGNO, M. (2004)
- CHOCIAI, R. (2004) Redação no vestibular da UNESP: a dissertação. São Paulo: Fundação VUNESP
- CHOMSKY, N. (1957) *Syntactic structures*. The Hague: Mouton&Co.
- _____ (1975) *Aspectos da teoria da sintaxe*. Coimbra: Arménio Amado. Trad. José António Meireles e Eduardo P. Raposo (texto original datado de 1965)
- DIAS-DA-SILVA, B. C. (1996) *A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais*. UNESP. Tese de doutorado.
- DUBOIS, J. et al (1993) *Dicionário de Lingüística*. São Paulo: Cultrix
- FARACO, C.A. (2004) Norma-padrão brasileira: desembaraçando alguns nós. In: BAGNO, M. (2004)

- FILLMORE, C. (1992) Corpus linguistics or computer corpus linguistics. In: SVARTVIK, J. (org.) *Directions in corpus linguistics. Proceedings of Nobel Symposium 82*, Stockholm, 4-8 August 1991, Berlin/NY: Mouton de Gruyter
- FLIP (2005) *Flip 5 – ferramentas para a língua portuguesa: Manual e Documentos*. Disponível apenas no site www.flip.pt (acesso e download em fevereiro de 2006).
- GOVINDANKUTTY, A. (1973) The computer and Dravidian linguistics. *ALLC Bulletin*, vol. 1, no. 3, p. 3-5.
- GRANGER, S. (2002) A bird's-eye view of learner corpus research. In: GRANGER, S., HUNG, J. e PETCH-TYSON, S. *Computer Learner Corpora. Second Language Acquisition and Foreign Language Teaching*. Amsterdam, Philadelphia: John Benjamins.
- HALLIDAY, M.A.K. (1991) Corpus studies and probabilistic grammar. In: AIJMER, K., ALTENBERG, B. (orgs.). *English corpus linguistics: studies in honour of Jan Svartvik*. Londres: Longman.
- HELFRICH, A., MUSIC, B. (2000) Design and evaluation of grammar checkers in multiple languages. *Proceedings of the 18th Conference on Computational Linguistics*, vol. 2. Saarbrücken, Germany, p. 1036-1040.
- HERDAN, G. (1964) Quantitative Linguistics or Generative Grammar. In: *Linguistics*, 4: 56-65.
- ISO/IEC 9126-1 (2001) *Standard – Software engineering – Product quality – Part 1: Quality model*. Geneva: International Organization for Standardization/International Electrotechnical Commission.
- JOSHI, A. K. (1991). Natural Language Processing. *Science*, Set., vol. 253, p. 1242-1249.
- KENNEDY, G. (1998). *An Introduction to Corpus Linguistics*. London/NY: Longman.
- KNOSHITA, J., SALVADOR, L.N., MENEZES, C.E.D. (2005) CoGrOO – um corretor gramatical para a língua portuguesa, acoplável ao OpenOffice. Disponível em http://cogroo.incubadora.fapesp.br/portal/down/Doc/Artigo_Clei_2005.pdf (Acesso e download em janeiro de 2006).
- KNUTSSON, O., PARGMAN, T. C., EKLUNDH, K. S. (2003) Transforming grammar checking technology into a learning environment for second language writing. *Proceedings of the HLT/NAACL 2003 workshop: Building Educational Applications Using NLP*. Edmonton, Canadá.
- LARA, G.M.P. (1994) A redação como tema de pesquisa. In: *Leitura: teoria & prática*, 13 (24): 66-82, nov.
- LEECH, G. (1991) The state of the art in corpus linguistics. In: AIJMER, K., ALTENBERG, B. (orgs.). *English corpus linguistics: studies in honour of Jan Svartvik*. Londres: Longman.
- _____ (1992) Corpora and theories of linguistic performance. In: SVARTVIK, J. (org.) *Directions in corpus linguistics. Proceedings of Nobel Symposium 82*, Stockholm, 4-8 August 1991, Berlin/NY: Mouton de Gruyter
- _____ (1997) Introducing corpus annotation. In: GARSIDE, R., LEECH, G., MCENERY, T. (orgs.) *Corpus annotation: linguistic information from computer text corpora*. Londres/New York: Longman.

- MARTINS, R.T. (2002) *PROBI: um corpus de teste para o revisor gramatical ReGra*. Relatório técnico NILC-TR-02-10. ICMC-USP.
- MATTOS E SILVA (2004) *O português são dois: novas fronteiras, velhos problemas*. São Paulo: Parábola
- McENERY, T., WILSON, A. (2001) *Corpus Linguistics: an introduction*. Edinburgh: Edinburgh University Press, second edition.
- MONTILHA, G; NUNES, M.G.V. (2000) *Testes Comparativos entre Revisores Gramaticais de Português*. Relatório técnico NILC-TR-00-9. ICMC-USP.
- MULLER, C. (1964) *Essai de statistique lexicale: l'illusion comique*. Paris: Klincksieck.
- NEVES, M.H.M. BORBA, F.S., LONGO, B.O., BAZZOLI, M. B., IGNÁCIO, S.E. (2005) *Dicionário Unesp do Português Contemporâneo*. 1. ed. São Paulo: Editora UNESP, 1.470 p.
- NUNES, M.G.V., MARTINS, R.T., HASEGAWA, R.; HABER, R.R.; MONTILHA, G. (2000) *Relatório dos Testes Comparativos entre Diferentes Versões do Revisor Gramatical ReGra*. Relatório técnico NILC-TR-00-8. ICMC-USP.
- NUNES, M.G.V., GHIRALDELO, C.M., MONTILHA, G., TURINE, M.A.S., OLIVEIRA, M.C.F., HASEGAWA, R., MARTINS, R.T., OLIVEIRA JR, O.N. (1996) Desenvolvimento de um sistema de revisão gramatical automática para o português do Brasil. *Anais do XIII Simpósio Brasileiro de Inteligência Artificial (SBIA'96)*. Curitiba, Brasil, 23-25 de outubro, p. 71-80.
- PARDO, T.A.S., RINO, L.H.M., NUNES, M.G.V. (2000) *Documentação e atualização do ReGra*. Relatório técnico, nº 109. ICMC-USP.
- PELIZZONI, J.M. (2002) *Preâmbulo ao aconselhamento ortográfico para o português do Brasil - Uma releitura baseada em utilidade e conhecimento lingüístico*. Dissertação de mestrado. ICMC-USP.
- PINHEIRO, G. M., ALUÍSIO, S. M. (2003). *Cópus Nilc: descrição e análise crítica com vistas ao projeto Lacio-Web*. Relatório técnico NILC-TR-03-03. ICMC-USP.
- RODRIGUES, A.D. (2004) Problemas relativos à descrição do português contemporâneo como língua padrão no Brasil. In: BAGNO, M. (2004).
- SANTOS, D. (2000) Introdução ao processamento de linguagem natural através das aplicações. In: RANCHHOD, E. (ed.) *Tratamento das línguas por computador. Uma introdução à lingüística computacional e suas aplicações*. Lisboa: Caminho, p. 229-59.
- SCHERRE, M.M.P. (2004) A norma do imperativo e o imperativo da norma: uma reflexão sociolingüística sobre o conceito de erro. In: BAGNO, M. (2004)
- SINCLAIR, J. M. (1991) *Corpus, Concordance, Collocation*. Oxford: OUP
- _____ (1992) The Automatic Analysis of Corpora. In: SVARTVIK, J. (org.) *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, Berlin/NY: Mouton de Gruyter
- _____ (1996) *EAGLES. Preliminary recommendations on Corpus Typology*. Disponível em http://www.ilc.pi.it/EAGLES96/corpus_typ/corpus_typ.html, acesso e download em dez/2004.

-
- _____ (2005) *Corpus and text: basic principles*. In: WYNNE, M. (org.) *Developing linguistic corpora: a guide to good practice*. Oxford: Osbow Books. Versão eletrônica disponível em <http://ahds.ac.uk/linguistic-corpora>. Acesso em out/2005.
- SMITH, G.W. (1991) *Computers and Human Language*. New York: Oxford.
- SOUZA, S.M.Z.L. (2001) *Avaliação e políticas educacionais: iniciativas em curso no Brasil. Educação e Estado: as mudanças nos sistemas de ensino do Brasil e Paraná na década de 90*. Londrina: UEL.
- SOUZA, S.M.Z.L., OLIVEIRA, R.P. (2003) *Políticas de avaliação e quase mercado no Brasil. Educação e Sociedade*, vol. 24, n. 84, set.
- TEMAA (1996) *TEMAA – Final report – A testbed study of evaluation methodologies: authoring aids. Technical Report LRE-62-070*. Center for Sprogteknologi, Copenhagen, Denmark.
- TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. Amsterdam/Philadelphia: Benjamins.
- VIANNA, H.M. (2000) *Avaliação educacional: teoria, planejamento, modelos*. São Paulo: Ibrasa.
- VIEIRA, R., LIMA, V. L. S. de. (2001) *Linguística computacional: princípios e aplicações*. Anais do Congresso da Sociedade Brasileira de Computação (SBC), 21. Fortaleza: SBC, v.2, p. 47-88.

Gisele Montilha Pinheiro

E-mail: gmontilha@usp.br
gmontilha@hotmail.com