Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL'2006), Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

Aquisição da Escrita Infantil: a Construção de um Corpus do Português Brasileiro

Thaïs Cristófaro-Silva[1], Raquel Fontes Martins[2], Daniela Oliveira Guimarães[3], Leonardo Almeida[4], Thiago Silva[5]

[1]thaiscas@hotmail.com, [2]raquel_fontesmartins@yahoo.com.br, [3]daniolive@yahoo.com, [4]sion701@gmail.com, [5]thfragasilva@yahoo.com.br

Abstract. This article describes the foundational aspects of the e-Labore project whose aim is to collect data from spontaneous texts written by children aged 6 to 12 years old. The corpus to be gathered will be made available to the academic community on www.projetoaspa.org/elabore. The children are from the city of Belo Horizonte, Brazil, and all study in private or state schools. A number of methodological aspects are discussed and evaluated critically. Preliminary results which reflect the first stage of this project are presented and issues for further enquiry are suggested.

Resumo. Este artigo descreve os principais aspectos do Projeto e-Labore cujo objetivo é coletar dados de textos escritos por crianças de 6 a 12 anos. O corpus a ser coletado será disponibilizado para a comunidade científica em www.projetoaspa.org/elabore. As crianças são da cidade de Belo Horizonte, Brasil, e todas estudam em escolas particulares ou públicas. Aspectos metodológicos são discutidos e avaliados criticamente. Resultados preliminares que refletem o primeiro estágio deste projeto são apresentados e são sugeridos tópicos para investigações futuras.

Keywords: Lingüística de corpus, aquisição da linguagem, linguagem e tecnologia

1 Introdução

Este artigo apresenta o "Projeto e-Labore: Laboratório Eletrônico de Oralidade e Escrita" e alguns de seus resultados iniciais. O Projeto e-Labore tem como objetivo central construir e disponibilizar para a comunidade científica um banco de dados de material escrito por crianças e pré-adolescentes de 6 a 12 anos, falantes do português brasileiro[1]. Essa faixa etária foi selecionada pois compreende o início da aquisição-aprendizado da escrita, a fixação e utilização do código escrito [3, 5, 6].

A iniciativa de se produzir um corpus escrito do português brasileiro infantil partiu, essencialmente, da observação da falta de corpora representativos nessa área. Corpora extraídos de livros didáticos elaborados para crianças são apresentados em [7] e [8].

[1] Lancaster Corpus Children's Project Writing (http://bowlandfiles.lancs.ac.uk/lever/index.htm) e Child Language Data Exchange System (http://childes.psy.cmu.edu/) são exemplos de corpora da escrita infantil, contudo, na língua inglesa.

Tais corpora são utilizados em pesquisas de aquisição da linguagem, contudo, eles não refletem diretamente a linguagem infantil, uma vez que são produções textuais formuladas por adultos. Há, também, um dicionário ilustrado elaborado para crianças a partir de estudo de corpora [1]. Contudo, os corpora utilizados na construção desse dicionário são também produções textuais de adulto.

O projeto e-Labore é uma iniciativa inovadora que pretende construir e disponibilizar para a comunidade científica um corpus contendo produções textuais espontâneas de crianças e pré-adolescentes. Um corpus dessa natureza poderá ser utilizado por pesquisadores de várias áreas do conhecimento: lingüistas, educadores, fonoaudiólogos, engenheiros da fala etc. Esses pesquisadores se beneficiarão de tal ferramenta para as mais diversas atividades de pesquisa e aplicabilidade tecnológica, tais como, projetos educacionais da língua portuguesa, estudos de patologias da fala infantil e formulação de teorias sobre a organização da linguagem humana.

O principal objetivo do Projeto e-Labore é possibilitar o conhecimento efetivo do léxico escrito de crianças e pré-adolescentes. A importância do conhecimento do léxico no aprendizado da leitura e da escrita é apontada em [2] e [4]. Tais estudos indicam a relevância e pertinência do presente projeto. O projeto pretende disponibilizar recursos importantes relacionados à linguagem escrita infantil, os quais podem oferecer contribuições para a investigação dos problemas atestados no processo de aquisição da escrita pelas crianças em idade escolar. A primeira investigação a ser empreendida pela equipe do projeto e-Labore concerne à formulação de uma taxonomia de desvios ortográficos a partir das produções escritas coletadas. Essa taxonomia deverá oferecer informações importantes sobre a superação de desvios ortográficos ao longo da vida escolar da criança e sua relação com a ampliação do léxico infantil.

Neste artigo, na Seção 2, apresenta-se a metodologia utilizada na construção do corpus do e-Labore, bem como, uma discussão a respeito da variabilidade na digitação das produções dos alunos. Em seguida, na Seção 3, discutem-se alguns resultados iniciais obtidos na primeira coleta piloto de redações. Por último, na Seção 4 é analisado o desenvolvimento futuro do projeto.

2 Metodologia

O corpus do projeto e-Labore é construído através da coleta e digitação de redações de crianças e pré-adolescentes com idade entre 6 e 12 anos que cursam da 1ª a 6ª série do ensino fundamental. O e-Labore conta com a colaboração de cerca de 15 alunos de graduação que são responsáveis por contatar as escolas, apresentar o projeto, instruir professores, coletar e digitar redações. A coordenação do e-Labore é formada pelos cinco autores deste artigo que orientam e supervisionam os estudantes de graduação em todas as etapas desenvolvidas. Há intenção de se realizar três coletas de redações: uma piloto e duas coletas completas. A coleta piloto, de menor escala, já foi realizada e, atualmente, todas as redações dessa coleta já foram digitadas, seguindo-se a metodologia descrita a seguir.

2.1 Coleta e organização das Redações

Todas as escolas participantes do e-Labore estão situadas na cidade de Belo Horizonte, Minas Gerais. A cidade de Belo Horizonte é dividida pela prefeitura local em 9 regionais: Barreiro, Centro-sul, Leste, Nordeste, Noroeste, Norte, Oeste, Pampulha, Venda-Nova. As escolas participantes do projeto estão uniformemente distribuídas de acordo com as regionais de Belo Horizonte. Sendo assim, em cada região da cidade 4 escolas - 2 da rede pública e 2 da rede particular - participam do projeto[2].

Ao coletar redações de alunos das redes pública e particular o projeto pretende avaliar se há diferenças no processo de aquisição da escrita entre os alunos desses dois tipos de escolas. O Projeto e-Labore não divulga dados específicos a respeito de nenhum aluno, professor ou escola participante. Deste modo, cada escola recebe uma carta do projeto na qual a coordenação deste se compromete a resguardar o anonimato de alunos, professores e da escola.

Os professores recebem uma folha contendo 8 instruções. As instruções procuram, principalmente, orientar o professor para que dê liberdade aos alunos durante a atividade de produção do texto. O professor também é instruído a não ajudar os alunos mesmo que solicitado. Além da folha de instrução, os professores recebem um questionário contendo perguntas sobre a experiência do professor e as dificuldades encontradas no ensino da língua portuguesa. As respostas deste questionário serão, futuramente, confrontadas com os resultados obtidos através da análise do corpus.

A coleta das redações é realizada como uma atividade de produção textual rotineira, pela professora da turma, na expectativa de não modificar as atividades regulares desenvolvidas pelos professores em sala de aula. O tema da redação em cada turma é definido pelos próprios professores e, sendo assim, há diversificação dos tópicos nas produções textuais. Essa decisão metodológica busca garantir que o vocabulário utilizado pelas crianças seja o mais próximo possível do que aquele que utilizam habitualmente em sala de aula.

Ao mesmo tempo em que coleta redações, o e-Labore também constrói um banco de dados contendo informações a respeito dos alunos e das escolas participantes do projeto. Em relação aos alunos, esse banco de dados contém informações a respeito do nome, data de nascimento, gênero e série. A cada aluno, é atribuído um número. Esse número é utilizado no acompanhamento longitudinal de sua produção de textos. Já em relação às escolas, os dados armazenados dizem respeito ao tipo de escola (pública ou particular), endereço, regional, telefone e pessoal responsável. O banco de dados relaciona informações a respeito de alunos e escolas às redações que compõem o corpus do projeto, atribuindo a cada produção um número de identificação único de 8 dígitos.

O número de identificação e um código de barras a ele associado são impressos em uma etiqueta e afixados ao papel contendo a redação. A utilização do código de barras permite que o número de 8 dígitos seja lido de maneira eficiente e rápida. O código de barras utilizado segue o padrão 2 de 5 entrelaçado. Esse código é o padrão da

---

2 É importante ressaltar, que na primeira coleta piloto apenas 14 escolas (2 particulares e 12 públicas) participaram do projeto. Nessa coleta o critério de distribuição das escolas pelas regionais da cidade não foi observado.

FEBRABAN (Federação Brasileira de Bancos), sendo largamente utilizado em boletos bancários e, desse modo, equipamentos que realizam a leitura de tal código de barras são facilmente encontrados no mercado.

Logo após serem identificadas, as redações têm suas imagens digitalizadas. A frente e o verso de cada uma das redações são escaneados. Sendo assim, toda a produção (textos, desenhos, palavras isoladas, acrósticos etc) realizada pelos alunos é registrada digitalmente. As imagens das redações são armazenadas em um arquivo no formato JPEG. Optou-se por escanear as redações com uma resolução bastante alta (3507x2480 pixels utilizando 24 bits por pixel). Essa resolução é capaz de garantir que nenhuma informação contida no original seja perdida. Além disso, pode-se, através de algoritmos bastante simples, reduzir essa resolução de tal modo que as imagens possam ser exibidas através da Internet. Todavia, antes de serem disponibilizadas na página do projeto, as imagens das redações deverão ser modificadas de modo a preservar totalmente o anonimato dos alunos.

2.2 Digitação das Redações

As redações produzidas pelos alunos são digitadas, uma a uma, pelos colaboradores do projeto. Muito embora a tarefa de digitação de redações pareça simples de ser realizada, na prática, encontram-se algumas dificuldades. Essas dificuldades decorrem não só do grande número de redações necessário para se construir um corpus representativo da linguagem utilizada pelos alunos em fase de aquisição da escrita, mas também da complexidade de se interpretar o desenho das letras de alunos que ainda atravessam o processo de alfabetização.

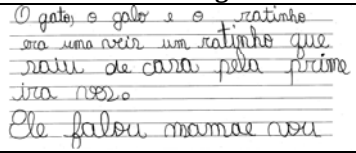
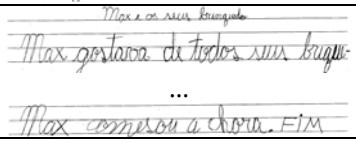
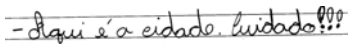
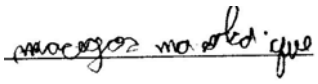
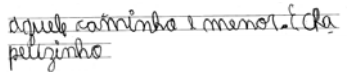
Na tentativa de padronizar o processo de digitação das redações, foi estabelecido um conjunto de 7 regras. As regras de digitação e exemplos aos quais elas se aplicam são apresentadas na Tabela 1. A regra 1 estabelece que os textos devam ser copiados de forma que a versão digitada seja a mais parecida possível com a versão original. Deste modo, as quebras de linhas e parágrafo realizadas pelos autores são apontadas na digitação das redações, preservando a organização espacial adotada na elaboração do texto.

A regra 2 estabelece que o caractere *$(cifrão)* marca o início e o fim do texto contínuo. O título da redação é digitado antes do primeiro cifrão que marca o início do texto. Por outro lado, falas em balão, poesias e fragmentos de texto fora do texto principal, são digitados após o cifrão que marca o fim do texto. Essa estratégia foi adotada na expectativa de se caracterizar o texto como unidade maior de sentido baseando-se estritamente no registro ortográfico.

Ao obedecer à regra 3, a versão digitada permitir a fácil identificação dos desvios ortográficos cometidos pelos alunos. Na presente etapa do projeto, optou-se por apontar apenas os desvios ortográficos. Sendo assim, desvios em colocações pronominais, regência (verbal ou nominal), concordância (verbal ou nominal) não são marcados. As palavras grafadas com desvios ortográficos pelos alunos são marcadas por um par de chaves e a forma padrão é apontada entre colchetes.

A regra 4 estabelece que a pontuação dos alunos deva ser integralmente copiada pelos digitadores. Sendo assim, mesmo que se encontrem erros claros no emprego da pontuação eles não são apontados.

Tabela 1. Exemplos de aplicação das regras de digitação.

N	Nome	Trecho Original	Trecho Digitado
1	Distribuição espacial		O gato, o galo e o ratinho era uma {veis}[vez] um ratinho que saiu de casa pela {prime ira}[primei_ra] {ves}[vez]. Ele falou {mamae}[mamãe] vou
2	Começo e fim de texto		Max e os seus brinquedo $Max gostava de todos seus {brique_ ... Max {comesou}[começou] a chora.$ FIM
3	Erro ortográfico		Uma {brucha}[bruxa] cheia de
4	Pontuação		- Aqui é a cidade. Cuidado!!!
5	Palavra ilegível		{macegos}[morcegos] {ma}[na] * que
6	Ausência de palavra		- Quem +[pega] o lixo é a Asmare.
7	Hifenização		aquele {e}[é] menor. E Cha_ peuzinho

Além dos símbolos utilizados para apontar desvios ortográficos, alguns outros foram criados para apontar diferentes situações. A regra 5 define que o símbolo **(asterisco)* dever ser empregado nas situações em que a palavra escrita é impossível de ser lida com precisão. Já a ausência de palavras que dão sentido ao texto, segundo a regra 6, é marcada por *+[palavra faltando]*.

Por último, a regra 7, estabelece que o caractere *"_"(underline)* deve ser utilizado na divisão de palavras em fim de linha. Deste modo, o caractere *"-"(travessão)* é utilizado exclusivamente, em palavras compostas e em marcação de fala de personagens. Essa decisão metodológica busca garantir a distinção clara entre palavras compostas ("bem-te-vi", "rouba-bandeira", "pisca-pisca" etc) e palavras divididas em sílabas devido ao término de uma linha (como a palavra "chapeuzinho" exemplificada na regra 7 da Tabela 1).

Todo o processo de digitação é realizado através da página do projeto na Internet. Os colaboradores recebem um nome de usuário e uma senha que são utilizados para acessar a interface de envio das redações. Os responsáveis pela digitação das redações passam por um processo de treinamento que busca garantir a uniformidade no processo de digitação. O treinamento começa com uma apresentação das regras. Em seguida, cada colaborador recebe 5 redações, acessa a interface de envio, e digita as

redações seguindo o conjunto de regras. Ao final do treinamento, a equipe de coordenação revisa as redações enviadas e discute com os colaboradores as inconsistências e erros cometidos na digitação.

Considerando-se que a digitação é realizada por um grande número de colaboradores, procura-se verificar a homogeneidade do processo através de um teste específico. A seguir, apresenta-se um teste que mede a concordância entre os colaboradores que participaram do processo de digitação de redações.

2.3 Teste de Concordância

Logo após definir-se a metodologia de digitação, fez-se necessário verificar a homogeneidade do cadastro das redações. O processo de digitação de uma redação pode ser considerado uma tarefa complexa. O colaborador deve obedecer ao conjunto de 7 regras de digitação discutidos na seção anterior e, ao mesmo tempo, interpretar corretamente a letra do autor da redação. Além disso, o colaborador está sujeito a cometer erros de digitação.

Com o objetivo de se avaliar a concordância dos digitadores com o padrão esperado pela coordenação do e-Labore, realizou-se um teste que contou com a participação de 10 colaboradores. Todos os colaboradores que participaram do teste digitaram mais de 100 redações. Juntos eles digitaram 1.434 redações, valor que corresponde a 73,5% do total de redações.

Para a realização do teste de concordância selecionou-se aleatoriamente 6 redações (uma de cada série). Todos os 10 colaboradores que participaram do teste digitaram essas 6 redações. Em média, os colaboradores gastaram 58 minutos para completar o teste.

Cada uma das redações digitadas no teste foi comparada com os padrões elaborados pela equipe de coordenação. Considerou-se que o colaborador discordou do padrão em cada uma das seguintes situações: palavras distintas, erro de pontuação, marcação incorreta de quebra de linha, marcação incorreta de parágrafo. No caso de palavras distintas, atribui-se apenas um erro por palavra. Somando-se todas essas situações chega-se ao número de 892 itens passíveis de erro.

A análise da Tabela 2 aponta que em média, o índice de concordância com o padrão é muito alto (96,1%). Pode-se notar, também, que o desempenho no processo

Tabela 2. Resultados do teste de concordância.

Série	Colaborador (concordância em %)										Média
	A	B	C	D	E	F	G	H	I	J	
1ª	95,7	95,7	96,8	88,2	93,5	86,0	95,7	93,5	86,0	84,9	**91,6**
2ª	89,6	89,6	91,0	89,6	89,6	91,0	97,0	91,0	95,5	82,1	**90,6**
3ª	98,2	98,2	99,4	97,6	97,0	96,4	97,6	95,8	98,8	97,6	**97,7**
4ª	96,0	93,4	97,4	97,4	98,0	95,4	99,3	97,4	98,0	96,7	**96,9**
5ª	97,5	98,6	97,9	96,4	96,8	97,9	97,5	96,4	95,7	95,7	**97,0**
6ª	97,8	97,0	98,5	97,8	97,8	95,5	99,3	97,8	97,8	94,0	**97,3**
Média	**96,6**	**96,4**	**97,5**	**95,6**	**96,3**	**95,1**	**97,9**	**96,0**	**96,0**	**93,8**	**96,1**

de digitação varia muito pouco de digitador para digitador. Um alto índice de concordância indica que os colaboradores foram capazes de entender e aplicar o conjunto de regras de digitação.

Por outro lado, a Tabela 2 indica que o desempenho dos colaboradores é pior na 1ª e 2ª série. Isso pode ser um indício de que uma grande parte dos erros cometidos na digitação das redações decorre da dificuldade de interpretação do desenho da letra do aluno. De fato, uma análise mais detalhada dos erros revela que 35,8% do total de erros advêm da dificuldade em se compreender a letra dos alunos.

Os erros cometidos na digitação das redações podem ser divididos em dois grupos. O primeiro grupo é composto por erros que não podem ser corrigidos em etapas futuras. Exemplos desse tipo de erro são palavras digitadas com grafia diferente da original, palavras esquecidas e marcação incorreta de linha ou parágrafo. O segundo grupo é composto por erros que podem ser corrigidos através de uma revisão manual. Nesse segundo grupo, pode-se incluir a digitação de palavras com erros ortográficos. Tais palavras tendem a aparecer com baixa freqüência de ocorrência na lista de palavras do corpus. Sendo assim, uma simples revisão nas palavras de baixa ocorrência pode eliminar esse tipo de erro do corpus.

Além da revisão manual das palavras, também se desenvolveu um algoritmo que é capaz de realizar uma revisão automática nas redações. Esse algoritmo verifica a correta utilização dos símbolos ($, {}, [ ], +[ ], _) nas regras 2, 3, 6 e 7. A aplicação desse algoritmo e a revisão manual das palavras de baixa freqüência de ocorrência reduzem em 37,3% o número de erros cometidos no teste. Deste modo, se fossem aplicadas tais correções às redações digitadas no teste, o índice geral de concordância com o padrão subiria para 97,6%.

3 Resultados Preliminares

O processo de digitalização e digitação da primeira coleta do projeto e-Labore está completo. Nesta seção, descreve-se o estado atual do corpus criado e apresentam-se, também, alguns resultados preliminares obtidos através da análise da lista de freqüência das palavras.

Na primeira fase, a equipe do projeto coletou 1952 redações em 67 turmas de 1ª a 6ª série, bem como digitalizou e digitou tais redações. A Tabela 3 apresenta o número de turmas, redações, palavras distintas (*types*) e total de palavras (*tokens*) que compõem o corpus ao final da primeira coleta. Atualmente, o corpus é formado por 11.415 palavras distintas que se repetem em um total de 207.459 palavras. Espera-se que, ao final das três coletas, o corpus do e-Labore atinja cerca de 1,5 milhão de palavras.

A contagem de freqüência de ocorrência das palavras no corpus levou em consideração apenas as palavras grafadas corretamente ou corrigidas pelos colaboradores. Assim sendo, desvios ortográficos não aparecem na lista de palavras distintas que ocorrem no corpus. Além disso, em busca de se obter um corpus livre de inconsistências, conferiu-se manualmente, uma a uma, as palavras que apresentam

Tabela 3. Situação do corpus do e-Labore após o encerramento da primeira coleta.

Série	Turmas	Redações	Palavras Distintas	Palavras
1a Série	12	324	2.285	22.476
2a Série	11	296	3.095	27.002
3a Série	12	379	4.137	42.582
4a Série	12	353	4.845	42.285
5a Série	10	296	4.319	37.659
6a Série	10	304	4.136	35.455
Total	**67**	**1.952**	**11.415**	**207.459**

Tabela 4. Lista de palavras mais freqüentes no corpus do e-Labore.

	Palavras			Substantivos			Verbos		
	Palav.	**Freq.**	**%**	**Subst.**	**Freq.**	**%**	**Verb.**	**Freq.**	**%**
1	e	9.622	4,64	natal	1.137	0,55	é	2.995	1,44
2	que	5.997	2,89	dia	1.108	0,53	foi	1.218	0,59
3	o	5.975	2,88	pessoas	993	0,48	tem	1.006	0,48
4	a	5.920	2,86	casa	747	0,36	era	833	0,40
5	de	5.528	2,66	mãe	589	0,28	estava	694	0,34
6	um	3.330	1,60	escola	510	0,25	ser	584	0,28
7	eu	3.307	1,60	ano	500	0,24	são	492	0,24
8	para	3.206	1,54	gente	496	0,24	tinha	491	0,24
9	não	3.058	1,47	mundo	446	0,22	está	481	0,23
10	é	2.993	1,44	projeto	384	0,19	vai	459	0,22
11	uma	2.268	1,09	crianças	381	0,18	fazer	429	0,21
12	com	1.969	0,95	família	327	0,16	ter	408	0,20
13	os	1.802	0,87	pai	323	0,16	vou	391	0,19
14	no	1.739	0,84	anos	323	0,16	pode	350	0,17
15	na	1.667	0,80	violência	315	0,15	acho	327	0,16

baixa freqüência de ocorrência. Esse procedimento busca excluir do corpus eventuais erros de digitação que possam ter sido cometidos pelos colaboradores.

A análise da lista de palavras do corpus e suas respectivas freqüências de ocorrência fornece resultados interessantes. A Tabela 4 apresenta as 15 palavras, os 15 substantivos e os 15 verbos mais freqüentes do corpus do e-Labore. Para cada um desses três casos, além da freqüência de ocorrência da palavra, apresenta-se também a porcentagem de ocorrência em relação a todo o corpus.

A respeito do primeiro caso, as 15 palavras mais freqüentes no geral, se pode observar que, excetuando-se *(9) não* e *(10) é*, todas as outras são palavras gramaticais e não palavras de conteúdo. Outra observação a ser feita é que as palavras desse primeiro caso são, em sua maioria, monossilábicas. Somente *(8) para* e *(11) uma* são dissílabos, contudo, tais palavras podem sofrer redução no discurso, passando a monossílabos: para → [ˈpa] ou [ˈpra] e uma → [ˈũə].

Sobre o segundo caso, os 15 substantivos mais freqüentes, deve ser ressaltado que eles, de modo geral, refletem os temas das redações propostos pelas professoras dos alunos participantes. Ainda, por exemplo, o substantivo *(1) Natal* reflete a época em

que as redações que compõem o corpus do e-Labore foram coletadas: final de ano. Muitas professoras sugeriram o Natal como tema das redações e, por esse motivo, tal substantivo aparece como o mais freqüente de todo o corpus. O caráter temporal da coleta em discussão, tendo como tema o Natal, poderá ser reavaliado a partir dos dados da segunda coleta que se dará no período pré-férias.

Quanto ao terceiro caso apresentado na tabela acima, os 15 verbos mais freqüentes, pode-se notar que, excetuando-se o verbo achar em *(15) acho*, todos os outros são verbos irregulares ("ser", "estar", "ir", "ter", "fazer" e "poder") [2]. Uma outra observação a ser feita quanto aos verbos mais freqüentes é que há 4 verbos no tempo passado: *(2) foi*, *(4) era*, *(5) estava* e *(8) tinha*. É provável que esses verbos no passado estejam refletindo o tipo textual utilizado na maioria das redações: a narração. Na narração, é muito comum o uso do tempo passado.

4 Conclusão

Este artigo descreve as principais características do Projeto e-Labore que tem como objetivo central construir e disponibilizar para a comunidade científica um banco de dados de material escrito por crianças e pré-adolescentes de 6 a 12 anos, falantes do português brasileiro. Inicialmente, apresenta-se a metodologia do projeto, discutindo aspectos a serem aprimorados em coletas futuras. Os principais pontos da coleta e da digitação do material escrito são apresentados, formulando uma perspectiva crítica que contribuiu para a coleta complementar de uma segunda etapa. Um teste de confiabilidade é discutido, uma vez que a digitação do corpus foi realizada por um grande número de participantes. Vale salientar que o conhecimento da modalidade escrita da faixa etária 6 a 12 anos do português é pouco explorado no português brasileiro. Esse conhecimento é importante, pois nessa faixa etária ocorre o início da aquisição-aprendizado da escrita, a fixação e utilização do código escrito, e um uso mais amplo da escrita na pré-adolescência. O conhecimento do vocabulário infantil nos moldes apresentados nesse projeto, permitirá a formulação e a disponibilização de uma ferramenta de pesquisa para várias áreas de investigação, tais como, projetos educacionais da língua portuguesa, estudos de patologias da fala infantil e formulação de teorias sobre a organização da linguagem humana. Os resultados finais do projeto devem oferecer contribuições importantes para o conhecimento do léxico na modalidade escrita de crianças e pré-adolescentes falantes do português brasileiro. Adicionalmente, os resultados desse projeto oferecerão a oportunidade de contraste entre o léxico infantil e adulto, o léxico de crianças com e sem patologias de fala e a investigação do desenvolvimento lexical escrito de crianças surdas (que utilizam gestos para a construção do léxico) e de crianças que utilizam a fala.

Agradecimentos. O projeto e-Labore conta com o financiamento do CNPQ processo 502906/2005-7.

Referências Bibliográficas

1. Biderman, M. T. C.: Dicionário Ilustrado do Português. Editora Ática (2005)

2. Bybee, J., Slobin D.: Rules and schemas in the development and use of the English past tense. Language 58 (1982) 265-89

3. Chambers, J.: Sociolinguistic Theory. Blackwell (1995)

4. Ellis, N. C.: Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition. SSLA, Cambrigde, 24 (2002) 143-188.

5. Labov, W.: Stages in the acquisition of standard English. In: Roger Shuy (ed.): Social Dialects and Language Learning. Ed. Champaign, Illinois (1964) 77-103.

6. Ministério da Educação.: Parâmetros Curriculares Nacionais. Documento disponível em http://www.mec.gov.br/sef/sef/pcn.shtm (2006)

7. Pinheiro, Â. M. V.: Leitura e escrita: contagem de freqüência de ocorrência e análise psicolingüística de palavras expostas a crianças na faixa pré-escolar e séries iniciais do 1° grau. Associação Brasileira de Dislexia, São Paulo (1996)

8. Pinheiro, G. M., Aluísio, S. M.: Corpus NILC: descrição e análise crítica com vistas ao projeto Lacio-Web. NILC-TR-03-03 (2003)