Capítulo 1

Os *corpora* orais e o C-ORAL-BRASIL

Heliana Mello

1. Introdução

A Linguística de *Corpus*,[1] como área disciplinar que explora *corpora* computadorizados, embora ainda tímida, tem gradativamente crescido no Brasil nas duas últimas décadas. Apesar de a tradição de estudos linguísticos baseada em dados reais da língua em uso ser muito mais antiga no país, por exemplo, nos estudos da Sociolinguística Variacionista (cf. Braga, 1977; Scherre, 1978),[2] e de muitos pesquisadores se referirem a seus bancos de dados não computadorizados como *corpora*, a compilação e disponibilização de *corpora* eletrônicos, sejam escritos ou orais, de fato existe no país há pouco tempo, certamente não o tempo suficiente para conferir aos linguistas brasileiros um número suficiente de recursos para que se possam efetivamente desenvolver estudos nos patamares daqueles que já vêm sendo desenvolvidos, principalmente na Inglaterra e nos países escandinavos, nos últimos 40 anos.[3] Nesses países, por exemplo, já há mais de três décadas vêm sendo publicados dicionários e gramáticas inteiramente produzidos a partir de grandes *corpora* de referência.

---

[1] Para um histórico completo da área, metodologias, assim como os seus subcampos de estudo, vejam-se Lüdeling; Kytö (2008, 2009) e referências seminais lá mencionadas.

[2] Cf. Hundt (2008), para uma discussão sobre as diferenças entre bancos de dados e *corpora*.

[3] Cf. "Histórico da Linguística de *Corpus*" (Berber Sardinha, 2004).

Desnecessário seria discutirmos as enormes vantagens oferecidas pelos *corpora* eletrônicos em relação às compilações de dados não eletrônicas que povoam os estudos linguísticos. O acesso público a dados linguísticos altamente sistematizados e computadorizados, assim como às ferramentas computacionais e estatísticas disponíveis para o seu tratamento, tornam hipóteses sobre a língua passíveis de testagem efetiva e redefinição, com altos níveis de representatividade e confiabilidade.

No cenário internacional, há *corpora* notáveis, como o British National Corpus (BNC),[4] com 100 milhões de palavras e passível de compra, ou ainda o Corpus of Contemporary American English,[5] compilado por Mark Davies, com 425 milhões de palavras e oferecido em uma plataforma que possibilita, gratuitamente, sofisticadas buscas *on-line*. Esses seriam *corpora* de referência, que pretendem representar uma dada língua como um todo; no caso, o inglês britânico e o estado-unidense, respectivamente. Há, entretanto, muitos outros tipos de *corpora*: a máxima, "um *corpus* vale de acordo com o seu objetivo" é inteiramente verdadeira. O valor de um *corpus* só pode ser medido em função de seu sucesso em atender os propósitos para os quais foi criado (cf. Lüdeling; Kytö, 2008, 2009). Assim, há *megacorpora* como os *corpora* monitores, que são frequentemente expandidos para acompanharem o percurso de uma dada língua (ex. Bank of English).[6] Há *corpora* diacrônicos e sincrônicos, de referência e especializados, *minicorpora*; há *corpora* criados para serem utilizados por décadas e outros para serem utilizados em um único projeto.

Berber Sardinha (2004, p. 9, 10) apresenta um quadro listando os principais *corpora* de língua portuguesa, tanto escritos quanto orais, e remete o leitor a outros autores que reportam projetos de desenvolvimento de bancos de dados do português. É notório que os estudos de *corpora* de língua portuguesa são mais bem estabelecidos em Portugal que no Brasil (cf. Bacelar do Nascimento *et al.*, 1996), já tendo-se firmado naquele país em centros de pesquisa, como o Centro de Linguística da Universidade de Lisboa, e em iniciativas, como a Linguateca,[7] notável portal criado pela linguista computacional Diana Santos.

---

[4] <http://www.natcorp.ox.ac.uk/>.

[5] <http://corpus.byu.edu/coca/>.

[6] <http://www.titania.bham.ac.uk/docs/svenguide.html>.

[7] <http://www.linguateca.pt/>.

No Brasil, como discutido por Berber Sardinha (2004, p. 6), as pesquisas baseadas em *corpora* são mais desenvolvidas por grupos envolvidos com o processamento da linguagem natural, Linguística Computacional e Lexicografia. Nesse sentido, os grupos de pesquisa com base no estado de São Paulo, como o NILC, o GELC, o COMET, dentre outros, têm sido pioneiros na compilação de *corpora* de grande importância para a Linguística de *Corpus* brasileira. Nos últimos anos, a pesquisa de *corpus* tem se expandido, como demonstrado pelas publicações e eventos científicos protagonizados pela área no Brasil.[8] A série de Encontros de Linguística de *Corpus* (ELC),[9] que ocorre anualmente, assim como a Escola Brasileira de Linguística Computacional são bons exemplos do fortalecimento desse campo disciplinar.

A maior parte dos *corpora* produzidos no Brasil são escritos, principalmente com material pertinente a jornais e gêneros acadêmico-científicos, utilizados, sobretudo, por grupos de pesquisa voltados para os estudos do léxico e desenvolvimento de ferramentas computacionais para o tratamento da linguagem natural; por exemplo, o *corpus* Lácio-Web.[10]

Berber Sardinha tem desenvolvido projetos de grande escopo, disponibilizando *corpora* e ferramentas de busca e tratamento estatístico. Exemplos desse trabalho seriam o *Corpus* Brasileiro, com um bilhão de palavras e disponibilizado *on-line*,[11] e o Banco do Português, também disponível parcialmente *on-line* e provido por uma série de ferramentas computacionais.[12]

A produção de *corpora* eletrônicos orais no Brasil, entretanto, ainda é bastante restrita e necessita ser fomentada. É nesse cenário que o C-ORAL-BRASIL foi vislumbrado e compilado. Seu objetivo principal é compor com o C-ORAL-ROM,[13] *corpus* europeu das quatro principais línguas românicas europeias, recurso linguístico computadorizado que permita estudos teóricos e aplicados da fala espontânea, com base empírica, pautada pela segmentação do fluxo sonoro em unidades formais correspondentes a unidades tonais. A cada segmento com quebra terminal corresponde um

---

[8] Cf. Berber Sardinha; Almeida (2008).

[9] Cf. <http://www.letras.ufmg.br/CMS/index.asp?pasta=linguisticacorpus2011&path=20101229104322.asp&title=Apresenta%E7%E3o&id=50>.

[10] <http://www.nilc.icmc.usp.br/lacioweb/corpora.htm>.

[11] <http://corpusbrasileiro.pucsp.br/x/>.

[12] <http://www2.lael.pucsp.br/corpora/bp/>.

[13] <http://lablita.dit.unifi.it/coralrom/>.

enunciado – que é a unidade fundamental de organização da fala (cf. Cresti, 2000a). Na seção 3, o C-ORAL-BRASIL e a sua caracterização serão mais pormenorizadamente apresentados.[14]

1.1 Os *corpora* orais no cenário internacional

No cenário internacional, a produção de *corpora* eletrônicos, marcada pela criação do Brown Corpus em 1964, um *corpus* de língua escrita, tem crescido a passos largos, seja em contexto acadêmico, empresarial ou em parcerias desses dois setores. Apesar de os *corpora* escritos ainda dominarem a produção na área, a compilação de *corpora* orais e multimodais tem se ampliado rapidamente. Os *corpora* orais têm encontrado crescente aplicabilidade não apenas nos estudos canônicos da Linguística (Sociolinguística, Dialetologia, Lexicografia, Morfossintaxe etc.), mas também no desenvolvimento de tecnologias da fala, tais como o reconhecimento e síntese da fala.

Bancos de dados de fala são conhecidos há mais tempo que os *corpora* da língua falada. A grande diferença entre esses dois construtos se deve ao fato de que, no primeiro, normalmente se fazem gravações estilizadas, baseadas em roteiros e efetuadas em cabines acústicas, com fins a estudos estritamente de cunho acústico e de aplicabilidade para a indústria da tecnologia da fala. Os *corpora* da língua falada, por outro lado, são normalmente baseados em desenhos específicos, para capturar a fala espontânea em suas várias modalidades. As grandes dificuldades associadas à gravação da fala espontânea e à sua representação – tais como necessidade de garantia de naturalidade da fala, limitações técnicas, nível de ruído, questões éticas ligadas à permissão para gravação, parâmetros para a transcrição, dentre outras – fizeram com que, durante anos, os *corpora* de fala se restringissem à *corpora* de conversas telefônicas, ou fala em contexto profissional, como palestras, aulas etc. (cf. Meyer, 2002).

Exemplos bem explorados dessa tipologia seriam o Switchboard[15] e o Corpus of Spoken Professional English.[16]

Há grandes *corpora* de referência, como o BNC acima mencionado, que incorporam *subcorpora* de fala. A parte oral do BNC, contudo, apresenta limitações, uma vez que não foi possível ser feito o seu balanceamento

---

[14] Veja-se o Capítulo 2 neste livro para uma caracterização minuciosa do C-ORAL-BRASIL.

[15] <http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html>.

[16] <http://www.athel.com/cpsa.html>.

integral (cf. Burnard, 2002), e há restrições severas na tipologia diafásica da amostra de fala interativa, já que a maior parte das gravações foi feita em ambiente familiar, durante as refeições, o que não permite uma variedade de tipos ilocucionários (cf. Moneglia, 2011).

A título de exemplificação, apresentaremos brevemente a seguir, dentre os projetos internacionais de grande representatividade para os estudos de *corpora* de fala, três casos que merecem especial destaque, por cumprirem a função de apresentarem *corpora* representativos da língua falada – compilados de acordo com padrões internacionais preestabelecidos, de forma a suprirem uma variedade tipológica representativa da fala –, por fazerem uso de anotação em diversos níveis e por serem disponibilizados através de seus arquivos de som, suas transcrições e o alinhamento de som e texto. São eles: o Santa Barbara Corpus of Spoken English, que hoje integra o International Corpus of English, o Spoken Dutch Corpus e o Reference Prosody Corpus of Spoken French.

O Santa Barbara Corpus of Spoken American English (SBCSAE)[17] é o componente norte-americano do International Corpus of English (ICE).[18] O ICE possui uma parte escrita e outra oral. Em sua parte oral, o ICE segue uma arquitetura específica observada por todas as variedades de inglês que o compõem, formada de 300 textos de cerca de 2 mil palavras cada. Esses, por sua vez, estão divididos em uma parte dialógica, com 180 textos, dos quais 100 são privados e 80 públicos; e outra monológica, com 120 textos, 70 dos quais não roteirizados e 50 roteirizados. O SBCSAE é dividido em quatro partes, compostas em sua totalidade por 60 segmentos discursivos. Os arquivos de som são disponibilizados pelo Linguistic Data Consortium, e as transcrições podem ser baixadas do *site* do projeto, que é licenciado pelo Creative Commons.[19] Os alinhamentos texto-som foram feitos através do *software* SoundWriter, que liga uma dada transcrição ao sinal sonoro a ela correspondente, além de permitir a visualização de parâmetros prosódicos da fala.[20]

---

[17] <http://www.linguistics.ucsb.edu/research/sbcorpus.html>.

[18] <http://ice-corpora.net/ice/index.htm>.

[19] <http://creativecommons.org/licenses/by-nd/3.0/us/>.

[20] <http://www.linguistics.ucsb.edu/projects/transcription/tools.html>.

O projeto original do Spoken Dutch Corpus (SDC)[21] previa um *corpus* de 10 milhões de palavras. Vários lançamentos de partes intermediárias desse *corpus* já foram feitos. No momento, ele conta com 9 milhões de palavras e cerca de 800 horas de gravação. Os arquivos de transcrição são conectados aos arquivos de som, através de um sistema próprio desenvolvido dentro do projeto, baseado em Hidden Markov Models. Parte da segmentação do *corpus* foi feita manualmente, e parte, automaticamente. Os transcritores e alinhadores utilizaram o *software* Praat[22] em seu processo de trabalho. Um *minicorpus* de cerca de 1 milhão de palavras foi transcrito foneticamente, havendo sido feita a validação do seu alinhamento com os arquivos de áudio. Excertos desse *corpus* contam com anotação sintática, assim como prosódica. Todo o *corpus* é lematizado e anotado morfologicamente. Adicionalmente, o *corpus* dispõe de uma interface, o COREX, que permite ao usuário realizar buscas cruzando todos os níveis de anotação do *corpus*.

O projeto francês Reference Prosody Corpus of Spoken French – Rhapsodie,[23] ainda em fase de construção, prevê a compilação de um *corpus* de referência do francês, representativo de diferentes gêneros discursivos, anotado prosódica e sintaticamente, com fins a amplos estudos da fala, que levem em conta a relação entre a prosódia, a sintaxe e a estrutura informacional. As unidades de segmentação da fala seguem os estudos de Claire Blanche-Benveniste (cf. Blanche-Benveniste, 2006), e os alinhamentos som-texto são feitos através do *software* WinPitch.[24]

1.2 Parâmetros para a compilação de *corpora* orais

Incluindo-se os princípios gerais que orientam a compilação de *corpora* – representatividade, amostragem e balanceamento –, os processos e procedimentos envolvidos na produção de *corpora* orais são motivo de debate na comunidade acadêmica. Alguns *corpora* orais são publicados apenas em sua transcrição (Wichmann, 2008),[25] o que a nosso ver impossibilita o real estudo da fala, uma vez que esta é, em grande parte, desprovida de sentido se não

---

[21] <http://lands.let.ru.nl/cgn/doc_English/topics/project/pro_info.htm#intro/>.

[22] <http://www.fon.hum.uva.nl/praat/>.

[23] <http://rhapsodie.risc.cnrs.fr/en/>.

[24] <http://www.winpitch.com/>.

[25] Cf. Sinclair; Mauranen (2006), que propõem um modelo sintático para o estudo da fala baseando-se em transcrições de *corpora* de fala e na capacidade interpretativa de quem executa a análise.

puder ser estudada em sua complexidade acústica (intensidade, duração, ritmo, timbre) associada ao nível léxico-morfossintático. Mesmo estudos da sintaxe da fala dependem crucialmente das unidades de referência em que é dividido o fluxo sonoro. Por exemplo, para uma sequência como (1) abaixo, adaptada do C-ORAL-BRASIL, desprovida de sua faceta acústica, em princípio poderiam ser propostas análises sintáticas muito distintas, baseadas em diferentes segmentações do encadeamento lexical ali presente:

1. aquele ali aquele ali já deu altura
   a) aquele // ali // aquele ali já deu altura //
   b) aquele ali // aquele // ali já deu altura //
   c) aquele ali // aquele ali já // deu altura //

A interpretabilidade de (1) está totalmente condicionada à sua contrapartida prosódica, ou seja, acústica. Nem mesmo sinais de pontuação ou qualquer outro tipo de notação seriam capazes de conduzir o pesquisador ao significado semântico, à estrutura sintática e às inferências pragmáticas de tal enunciado. Apenas a oitiva acompanhada da transcrição devidamente executada podem, em conjunto, servir de instrumento para a adequada análise linguística de tal produção.[26]

A compilação de um *corpus* oral deve partir das perguntas iniciais: para quê e para quem ele servirá? Uma vez que haja um perfil de aplicabilidade para o *corpus*, é necessário que se definam os seus parâmetros de compilação. Estes estão associados a fatores como: tamanho; número e perfil dos informantes; técnicas para coleta/eliciação de dados; conteúdo e desenho do *corpus*: gênero (conversação, diálogo, monólogo); variação linguística (geográfica, contextual, individual); tipologia e qualidade dos arquivos de som; unidades de referência da fala, parâmetros de transcrição, acessibilidade; alinhamento texto-som; níveis de anotação; validação; custo; disponibilização (cf. Wichmann, 2008).

Todos os fatores acima devem ser considerados e ajustados aos propósitos a que o *corpus* servirá, definindo-se os padrões a serem adotados. Há pontos relacionados aos parâmetros norteadores da compilação de *corpora* orais que devem ser considerados cuidadosamente:[27]

---

[26] Cf. Capítulo 2 neste livro para uma discussão detalhada sobre este assunto; Cf. Moneglia (2011) para uma minuciosa discussão da inseparabilidade entre arquivos de som e de transcrição em *corpora* orais e a análise informacional da fala.

[27] Cf. Oostdijk *et al.* (2002) para uma discussão sobre a experiência do Spoken Dutch Corpus Project.

a) questões legais e éticas relacionadas à identidade dos informantes – seja por reconhecimento de nomes ou voz – e das informações veiculadas em suas falas. É necessário que os informantes assinem um termo de consentimento através do qual fiquem estabelecidos os elementos autorizados a serem divulgados ou não;

b) equipamentos e *software* a serem utilizados para a gravação e tratamento dos arquivos de som;

c) treinamento da equipe de trabalho para a execução de todas as etapas do projeto;

d) técnicas de validação e tratamento estatístico dos dados;

e) interface de acesso ao *corpus*.

Como se pode inferir, o pré-planejamento do desenho do *corpus*, a clara definição dos parâmetros a serem adotados e uma grande capacidade de gerenciamento de seus executores são todos pré-requisitos para a eficiente compilação de um *corpus* que possa atender aos seus propósitos.

A seguir, na seção 2, serão apresentados alguns bancos de dados e *corpora* orais brasileiros amplamente conhecidos, objetivando-se oferecer uma panorâmica dos recursos disponibilizados aos pesquisadores na área, sem a pretensão de se fazer uma apresentação exaustiva dos mesmos.

2. Os bancos de dados e *corpora* orais brasileiros

A própria natureza dos dados linguísticos orais dificulta a sua compilação em *corpora*. Enquanto dados escritos em formato eletrônico podem, literalmente, ser incorporados automaticamente a um dado *corpus*, como comprovado pela compilação de *corpora* extraídos diretamente da *web*, via *software* como o BootCaT,[28] ou mesmo a utilização da própria *web* como *corpus* assim como é feito pela WebCorp,[29] os dados orais carecem de tratamento muito minucioso e específico na sua captura, após um complexo período de arquitetura e definição do *corpus* que virão a compor. Primeiramente, a captura dos dados deve ser feita através de equipamento de gravação sofisticado o suficiente para preservar a boa qualidade de som e a naturalidade da fala, o que posteriormente permitirá a transcrição de dados com maior índice de acuidade e, consequentemente, a produção de

[28] <http://bootcat.sslmit.unibo.it/>.

[29] <http://www.webcorp.org.uk/live/>.

estudos baseados em análises apuradas das qualidades acústicas da onda sonora. Em segundo lugar, a fala espontânea raramente se dá em forma estritamente monológica. No universo da existência humana, a fala se presta muito mais à interação dialógica e multilógica, o que acarreta a sobreposição de turnos, interferências, falsos começos etc. e, consequentemente, sua maior dificuldade de representação em forma ortográfica.

As decisões relacionadas ao código de transcrição da fala não são triviais e exigem dos pesquisadores um grande esforço metodológico para que se atinjam soluções satisfatórias, que cumpram o papel de explicitar fenômenos característicos da fala, sem contudo dificultar ao extremo a leitura dos dados pelo usuário do *corpus*. A representação da fala deve trazer, além da representação de fenômenos léxico-gramaticais característicos dessa diamesia,[30] a segmentação do encadeamento sonoro em unidades de referência compatíveis com a organização da fala. Não se pode pensar em uma segmentação para a fala baseada em sentenças ou predicações; é necessário que a intenção comunicativa dos informantes seja bem representada, isto é, que as suas ações verbais sejam preservadas.[31] A fala organiza-se em enunciados, subdivididos em unidades tonais, e este encadeamento deve ser mantido na codificação transcritória do sinal sonoro.

Uma outra questão relacionada ao processo de transcrição da fala diz respeito ao treinamento dos transcritores, que exige tempo e consistência, para que se forme uma equipe com altos níveis de competência e de concordância. Para tal, é necessário um grande empenho na capacitação da equipe e a certificação dos resultados do seu trabalho através de testagem e validações estatísticas sobre o grau de convergência de suas decisões de transcrição, seja no nível segmental ou suprassegmental.

Um terceiro complicador é a necessidade de alinhamento entre os arquivos de áudio e as suas transcrições, a fim de que o usuário possa observar os dados a serem estudados, através do cotejamento entre o sinal sonoro, suas propriedades acústicas e sua transcrição. Caso esse estágio não seja observado na compilação de *corpora* orais, torna-se infactível a realização de uma série de estudos de caráter fonético-fonológico, semântico, sintático e pragmático que dependem da análise minuciosa do sinal sonoro e suas características constituintes em cotejamento com a sua

---

[30] Cf. Berruto (1993).

[31] Veja-se o Capítulo 3 neste livro para uma minuciosa discussão sobre a noção de enunciado e da organização informacional da fala.

transcrição. É literalmente impossível para um pesquisador, concomitantemente, ouvir arquivos de som, acompanhar a sua transcrição e formular uma análise dos fenômenos de seu interesse *in abstracto*. A análise de dados da fala através de suas transcrições apenas, como efetuado em muitos projetos, reduz, literalmente, a fala a uma codificação escrita, destituindo-a de suas características próprias. Até mesmo estudos lexicográficos ficam comprometidos sem a adequada representação som-transcrição da fala. O encadeamento entre sinal sonoro e representação gráfica deve ser oferecido como elemento constitutivo de um *corpus* oral.

Some-se às dificuldades já listadas o fato de *corpora* orais dependerem da autorização dos falantes para que sejam gravados. Faz-se necessário que se encontrem indivíduos, nas mais diversas situações interacionais, que permitam a gravação de suas falas e ações linguísticas, já não se admitindo mais contemporaneamente a gravação secreta de informantes. O desenvolvimento de tecnologias de gravação tem nos possibilitado capturar com acuidade e espontaneidade a fala, por meio de microfones sem fio, com transmissão de sinal via rádio e gravação em vários canais, via *mixer*, para interações de dois ou mais indivíduos.

Um *corpus* oral que pretenda representar a fala espontânea deve contemplar as diferentes tipologias do discurso oral, de forma balanceada e representativa. Isso leva a uma outra barreira, que é a do tempo. O tempo de produção de um *corpus* oral é necessariamente longo, depende de uma equipe bem articulada e treinada, e isso nem sempre foi apreciado pelas instituições de suporte à pesquisa. Dadas as dificuldades listadas, uma consequência inevitável é o custo de produção de *corpora* orais, seja ele estritamente financeiro ou ligado ao empenho de tempo e qualificação da equipe, que claramente é muito superior àquele de *corpora* escritos. Os empecilhos que se colocam à criação de *corpora* orais são reais e desestimulam os pesquisadores da área; isso pode ser facilmente comprovado observando-se o índice de projetos propostos e iniciados na área, mas nunca efetivamente concluídos. Entretanto, as equipes que perseveram em suas metas de compilação de *corpora* orais, apesar das dificuldades, têm nos oferecido produtos que possibilitam que estudos da fala possam ser efetuados cada vez com maior amplitude de campos analíticos e precisão metodológica, levando adiante a pesquisa e a compreensão da estrutura da fala.

A seguir serão listados e brevemente descritos os principais projetos brasileiros que de alguma forma compilam ou utilizam banco de dados e

corpora orais. Esses repositórios, nascidos normalmente em projetos que não estão direcionados à compilação de *corpora per se*, visam a estudos teoricamente orientados, frequentemente estudos da Variação Linguística ou do Discurso, voltados para aspectos particulares da língua falada, como o léxico e a sintaxe (ou, às vezes, em combinação com o estudo da escrita), e não foram, portanto, concebidos para a compreensão da organização da fala como um sistema. Os projetos listados mantêm algum nível de informação sobre si, via acesso pela *web*, mas, de modo geral, as informações são escassas, não havendo detalhamentos sobre o processo de compilação, transcrição e armazenamento dos *corpora*. Ver-se-á que há uma variedade de arquiteturas, propósitos, assim como variados níveis de acessibilidade e documentação, temas a serem retomados na seção 3 deste capítulo, quando o C-ORAL-BRASIL será discutido tendo em vista o cenário de *corpora* orais no Brasil.

2.1 NURC – Projeto Norma Linguística Urbana Culta

O Projeto Norma Linguística Urbana Culta, iniciado em 1970, foi precursor no Brasil e talvez seja a mais conhecida iniciativa brasileira no campo dos estudos da língua falada. Surgiu através da adesão, viabilizada por Nelson Rossi, da UFBA, em 1968, ao Proyecto para el Estudio de la Norma Linguística Culta, elaborado por Juan M. Lope Blanch, em 1967, como relatado por Castilho (2005). Tal iniciativa agregou grupos de pesquisadores de diversas procedências no Brasil, nomeadamente das cidades de Recife, Salvador, Rio de Janeiro, São Paulo e Porto Alegre. A amostra coletada nessas cinco cidades seria indicativa da norma da fala urbana culta do Brasil em virtude de elas agruparem aproximadamente um sétimo dos habitantes do país em áreas urbanas antigas, já que, exceto por Porto Alegre, fundada no século XVIII, as demais foram fundadas no século XVI.

O projeto NURC produziu uma extensa lista de publicações, ampliada posteriormente pela utilização dos seus dados no Projeto de Gramática do Português Falado nos anos de 1980 e 1990 (cf. Castilho, 2005).

Os dados gravados de forma magnetofônica pelo projeto NURC não estão disponibilizados *on-line* para a comunidade, exceto parcialmente pelo projeto NURC-RJ (cf. 2.1.2), sendo que os dados coletados pelas demais equipes do projeto têm acesso restrito. As transcrições das gravações estão disponibilizadas através de publicações impressas.

Como descrito pelas equipes NURC-RJ e NURC-Salvador,[32] o projeto NURC conta com os mesmos princípios metodológicos nas cinco cidades, quais sejam: (a) informantes dos dois gêneros, agrupados em três faixas etárias: 25 a 35 anos, 36 a 55 anos e 56 anos em diante. Os informantes devem ter nascido e permanecido nas cidades em estudo pelo projeto por pelo menos três quartos de sua vida; (b) categorias textuais compostas por elocuções formais-EF (aulas e conferências), diálogos entre informante e documentador-DID e diálogos entre dois informantes-D2. O conjunto de dados do *corpus* NURC, em sua maioria gravado na década de 1970, perfaz um total de 1.870 inquéritos gravados, correspondentes a 1.570 horas de gravação. Tal conjunto amplia-se com a adição de novas gravações efetuadas por diferentes equipes de trabalho.

Dados os seus propósitos (estudo da norma urbana culta), o projeto NURC adotou critérios de transcrição ortográficos, com marcações específicas para ênfase, truncamento, pausa etc. Não há marcação de pontuação nem de qualquer aspecto morfossintático ou pragmático da fala, assim como não há menção a qual seria a unidade de referência da fala (cf. Preti; Urbano, 1990). A título de exemplificação, veja-se abaixo um excerto de transcrição do NURC-RJ:[33]

(1)

DIÁLOGOS ENTRE DOIS INFORMANTES (D2)

Inquérito 296

L1: mas escuta uma coisa... que que ele diz dessa situação... desse centro da cidade aqui ( ) nem varrem mais essa rua ( )

L2: eles tão eles tão reorganizando essas essas ruas... porque eles querem acabar aqui no centro da cidade com o estacionamento de carros... porque eles acham que isto prejudica muito o tráfego e na hora do "rush" fica muito atravancado então... eles querem dificultar o mais possível... que os particulares tragam os seus carros pra cidade ( ) com o estacionamento...

---

[32] <http://www.twiki.ufba.br/twiki/bin/view/Alib/AlibNurc> e <http://www.letras.ufrj.br/nurc-rj/>. Acessos em: 14 out. 2011.

[33] <http://www.letras.ufrj.br/nurc-rj/>. Acesso em: 14 out. 2011.

A seguir, passamos a uma breve apresentação dos dados disponibilizados pelas equipes que compõem o projeto NURC.

2.1.1 NURC-SP

O projeto NURC-SP foi coordenado de 1969 a 1980 por Ataliba Castilho e Isaac Nicolau Salum, posteriormente por Ataliba Castilho e Dino Preti, e é atualmente coordenado por Dino Preti e Marli Leite. O braço paulistano foi protagonista na história do projeto NURC e gerou importantes iniciativas de sistematização e análise dos dados coletados no projeto em todas as cidades parceiras. Hoje, constitui-se em grupo de pesquisa cadastrado na Plataforma do CNPq[34] e USP.[35]

A vertente paulistana do NURC não disponibiliza o seu *corpus* através da *web*, porém, como já mencionado, a sua impressionante lista de publicações traz, além de estudos diversos, as transcrições de dados.

2.1.2 NURC-RJ

Coordenado por Dinah Callou, o projeto NURC-RJ é o único do país que disponibiliza uma grande quantidade de seus dados orais e transcritos em seu *site*.[36] Os trabalhos do projeto estão também disponibilizados em uma longa lista de publicações, incluindo-se as transcrições das gravações (cf. Callou, 1992; Callou; Lopes, 1993; 1994). De acordo com as informações pormenorizadamente disponibilizadas em seu *site*, ao longo de sua história, o projeto tem atuado em três subprojetos, quais sejam, Fonética/Fonologia, Morfossintaxe e Léxico. O subprojeto de Fonética e Fonologia, partindo da Sociolinguística Laboviana e da Fonética Experimental, concentrou-se em: a análise articulatória e acústica do vocalismo átono e tônico; o processo de harmonia vocálica; a análise das consoantes que travam sílaba (*s*, *r*, *l*); a nasalização, a ditongação e a palatalização.

Na vertente morfossintática do projeto, os pesquisadores, adotando preceitos variacionistas e funcionalistas, já exploraram uma vasta gama de temas, incluindo-se o sistema verbal e seus processos de concordância,

---

[34] <http://dgp.cnpq.br/buscaoperacional/detalhegrupo.jsp?grupo=00678013FPW609>.

[35] <https://sistemas.usp.br/tycho/gruposPesquisaObter?codigoGrupoPesquisa=0067801U3BKFOV>.

[36] <http://www.letras.ufrj.br/nurc-rj/>.

o sistema pronominal, indeterminação de sujeito, aspectos discursivos variados, dentre outros.

Os estudos do léxico desenvolvidos pela equipe enfocaram substantivos, adjetivos e verbos. Desenvolveram-se duas listas lexicais, uma de itens léxico-textuais e outra de lexemas, acompanhados de sua frequência absoluta.

A equipe NURC-RJ tem dado prosseguimento ao seu trabalho e relata novas gravações efetuadas na década de 1990, com o intuito de ampliar o seu *corpus* e de coletar material direcionado a estudos voltados à mudança linguística.

2.1.3 NURC-Salvador

Há escassas informações sobre o projeto NURC-Salvador em seu *site*.[37] Os dados do projeto em Salvador foram coletados dentro das mesmas tipologia e metodologia anteriormente descritas (seção 1.1). Na primeira fase do projeto, coletaram-se 307 horas e 20 minutos de gravação, documentando 461 informantes distribuídos em 360 inquéritos. As gravações foram efetuadas em fitas de rolo e em fitas cassete, dada a tecnologia disponível na época. Deu-se prosseguimento à coleta de dados na década de 1990. Os dados de áudio do projeto em suas versões originais e em formato digital encontram-se armazenados no Arquivo Sonoro do Setor de Língua Portuguesa do Instituto de Letras da Universidade Federal da Bahia. Assim como as demais equipes, o NURC-Salvador produziu uma vasta gama de trabalhos científicos (cf. Motta; Rollemberg, 1994).

2.1.4 NURC-Recife

O *corpus* coletado pela equipe NURC-Recife não se encontra disponibilizado no *site* do projeto.[38] As informações lá listadas relatam a coordenação inicial do projeto em 1971 por José Brasileiro Tenório Vilanova. Atualmente, o projeto se insere na linha de pesquisa "Análise Sociopragmática do Discurso", do Programa de Pós-Graduação em Letras da Universidade Federal de Pernambuco. Dados do projeto foram publicados em Sá *et al.* (1996).

---

[37] <http://www.twiki.ufba.br/twiki/bin/view/Alib/AlibNurc>.

[38] <http://www.pgletras.com.br/programa-nucleos-nurc.htm>.

2.1.5 NURC-Porto Alegre

O projeto NURC-POA não dispõe de *site* próprio. Suas publicações estão disponibilizadas em forma impressa (cf. Hilgert, 1997).

2.2 Projeto Discurso e Gramática

O Projeto Discurso e Gramática, sediado na UFRJ,[39] obteve seu primeiro projeto integrado, denominado Iconicidade na fala e na escrita, em 1991. A produção do *corpus*, com uma componente escrita e outra falada, possui amostras das cidades do Rio de Janeiro, Natal, Rio Grande, Juiz de Fora e Niterói. O projeto disponibiliza parte de seus dados em seu *site* (subdivididos nos gêneros: narrativa de experiência pessoal, narrativa recontada, descrição de local, relato de procedimento e relato de opinião) e lá apresenta os seguintes objetivos de pesquisa, aqui citados literalmente:

a) analisar o comportamento da iconicidade, através de diferentes fenômenos linguísticos, em situações reais de uso da língua;

b) criar um banco de dados com correspondência de conteúdo entre fala e escrita, de modo a viabilizar a comparação mais rigorosa entre essas duas modalidades da língua;

c) testar em diferentes subgêneros textuais (narrativa de experiência pessoal, narrativa recontada, descrição de local, relato de procedimento e relato de opinião) o modo de codificação da informação;

d) comparar o comportamento dos canais da fala e da escrita em relação a esses subgêneros.

Não há informações disponíveis sobre o processo e os critérios de transcrição da fala adotados por esse projeto.

2.3 VARPORT: Análise Contrastiva de Variedades do Português

Este projeto, também sediado na UFRJ, é uma iniciativa que vincula pesquisadores brasileiros do Setor de Língua Portuguesa do Departamento de Letras Vernáculas da Universidade Federal do Rio de Janeiro e pesquisadores portugueses do Centro de Linguística da Universidade de Lisboa (CLUL).[40] Os *corpora* que integram o projeto são vários e compostos tanto

---

[39] <http://www.discursoegramatica.letras.ufrj.br/>.

[40] O *site* do projeto, contendo extensas informações descritivas sobre os seus propósitos, é: <http://www.letras.ufrj.br/varport/>.

de língua escrita quanto falada e serão listados abaixo. Alguns deles estão disponibilizados, pelo menos parcialmente, através do *site* do CLUL,[41] assim como também o está a vertende carioca do NURC, como já explicitado na subseção 2.1.2. Os temas específicos da pesquisa comparativa do projeto são listados como a seguir:

> No âmbito de cada uma das variedades do Português, descrever no plano fonético fonológico: os padrões prosódicos que singularizam as variedades nacionais, a atuação de processos de enfraquecimento vocálico e consonantal, a nasalidade vocálica. No plano morfossintático descrever: ordem dos constituintes no nível da oração e da frase; sujeito preenchido e sujeito nulo, topicalização, verbos leves e verbos plenos, regência nominal e verbal, dêiticos, processos morfológicos flexionais e derivacionais. No plano morfológico, morfofonológico e morfossintático, descrever: padrões de flexão verbal, concordância SUJ-V, formas nominais do verbo. No plano léxico-semântico descrever: a frequência e distribuição de itens lexicais.[42]

Os *corpora* que integram o projeto são os seguintes: na componente portuguesa, estão disponíveis o *Corpus* do Português Fundamental, o *Corpus* de Referência do Português Contemporâneo (CRPC) e as elocuções livres do *Corpus* do Atlas Linguístico-Etnográfico de Portugal e da Galiza (ALEPG), do Centro de Linguística da Universidade de Lisboa. Na componente brasileira, integram o projeto o Arquivo Sonoro do Projeto Norma Urbana Culta (NURC), o Arquivo Sonoro do Projeto do Atlas Etnolinguístico dos Pescadores do Estado do Rio de Janeiro (APERJ) e o *Corpus* do Português Clássico e Moderno, no qual se inclui o *Corpus* do Brasil Colônia.

O Projeto NURC foi caracterizado na seção 2.1. Já o Projeto do Atlas Etnolinguístico dos Pescadores do Estado do Rio de Janeiro (APERJ) conta com um *corpus* de 178 horas de gravação no Norte-Noroeste do estado do Rio de Janeiro, correspondente a entrevistas com 78 informantes, realizadas em 13 localidades daquela região. Esse *corpus* ainda está em construção, e pretende-se expandi-lo para a inclusão das regiões das Lagunas Litorâneas, Metropolitana e Sul do estado do Rio de Janeiro, prevendo-se a inclusão de 36 outras comunidades. O *Corpus* do Português Clássico e Moderno, ainda em processo de compilação, agrupa produção manuscrita no Brasil durante

---

[41] <http://www.clul.ul.pt/pt/recursos>.

[42] Citação extraída da seção de objetivos do Projeto VARPORT. Disponível em: <http://www.letras.ufrj.br/varport>. Acesso em: 1º dez. 2011.

o Período Clássico (séculos XVI, XVII e XVIII) e produção manuscrita e impressa no Brasil durante o Período Moderno (séculos XIX e XX).

Como se nota, o projeto VARPORT tem um caráter amplo e seus *corpora* de estudo são bastante heterogêneos em sua composição, propósitos e, naturalmente, seus critérios de transcrição e anotação.

2.4 PROFALA: Variação e Processamento da Fala e do Discurso: análises e aplicações

De acordo com o *site* do projeto PROFALA, sediado na Universidade Federal do Ceará, o seu objetivo geral é "a implantação de um sistema baseado em tecnologia da informação para análises e aplicações à língua falada e ao discurso".[43] O projeto visa ao desenvolvimento de análises linguísticas (fonético-fonológicas, léxicas, morfossintáticas, pragmáticas, discursivas), dialetais (variações diatópicas do falar do Ceará e de outros estados nordestinos), sociolinguísticas (variações diastráticas do falar do Ceará e de outros estados nordestinos) e psicolinguísticas (processamento da fala e do discurso). O projeto tem à sua disposição os seguintes *corpora* já existentes na Universidade Federal do Ceará: o Português Não Padrão do Ceará, o Português Oral Culto de Fortaleza, Projeto AliB-CE (cf. Aragão; Soares, 1996). Há uma pequena amostra de três arquivos de som, disponíveis no *site* do projeto. O *corpus* do projeto PROFALA foi coletado em cidades da região do Cariri, especialmente Barbalha, Nova Olinda, Juazeiro, Várzea Alegre, Altaneira, Mauriti, Caririaçu e Brejo Santo. As transcrições das entrevistas estão disponíveis no *site* do projeto.[44] Não há menção aos critérios de transcrição adotados para tal.

2.5 VALPB: Projeto Variação Linguística no Estado da Paraíba

O projeto VALPB exibe apenas uma explicação genérica sobre os seus propósitos em seu *site*.[45] Criado em 1993, o projeto objetiva o estudo da variação sociolinguística da fala de João Pessoa. Foram efetuadas publicações baseadas no *corpus*, o qual encontra-se digitalizado e também impresso, porém não disponível no *site* (cf. Hora; Pedrosa, 2001).

---

[43] <http://www.profala.ufc.br/>.

[44] <http://www.profala.ufc.br/tabela1.htm>.

[45] <http://ibraed.com/index.php?option=com_content&view=section&layout=blog&id=6&Itemid=61>.

2.6 Projeto Vertentes do Português Popular do Estado da Bahia

O projeto Vertentes[46] objetiva traçar um panorama sociolinguístico do português popular falado do estado da Bahia. Tal propósito está calcado no estudo de comunidades rurais afro-brasileiras isoladas, assim como da fala da cidade de Salvador. Os dados de fala do projeto concentram-se na norma popular.

O projeto adota o enquadramento teórico-metodológico da Sociolinguística Variacionista, com aportes da Teoria da Gramática Gerativa, para realizar análises dos tópicos mais significativos da morfossintaxe do português popular brasileiro, através das quais busca diagnosticar os processos de variação estável e mudanças em progresso que definem as tendências atuais da língua no Brasil, por um lado, e identificar os reflexos do contato entre línguas presente na formação histórica das variedades linguísticas analisadas, por outro. O *corpus* não é disponibilizado para o uso da comunidade acadêmica, restringindo-se aos membros do projeto. As diretrizes adotadas para a transcrição da fala estão explicitadas no *site* do projeto e deixam claro o objetivo de privilegiar fenômenos de caráter morfossintático, como pode ser visto no excerto a seguir, que justifica a opção por se adotar a pontuação canônica, como marca de fronteira de fenômenos sintáticos: "uma transcrição que se funda nessa sinalização da estrutura sintática do texto oral revelou-se a mais eficaz para conferir inteligibilidade ao texto transcrito e a mais profícua para a análise que se tem realizado no âmbito do projeto Vertentes."[47]

O projeto Vertentes contou com uma série de publicações ao longo dos seus 15 anos de existência, sendo a mais notável a obra de Lucchesi; Baxter; Ribeiro (2009).

2.7 VARSUL: Variação Linguística na Região Sul do Brasil

O projeto VARSUL[48] foi criado como um consórcio de equipes de três universidades federais do Sul do Brasil (UFRGS, UFSC e UFPR), em 1982, seguindo a proposta de Leda Bisol de se organizar um banco de dados linguísticos da região Sul do país. Quando da sua proposição, os objetivos

---

[46] <http://www.vertentes.ufba.br/home>.

[47] Cf. <http://www.vertentes.ufba.br/projeto/transcricao>.

[48] <http://www.varsul.org.br>.

estavam voltados aos seguintes temas de pesquisa: constituição de um atlas linguístico e etnográfico, bilinguismo e variação linguística. Com a colaboração das equipes e agregação de novos membros, obtenção de fomento e criação de planos de trabalho, o projeto VARSUL, hoje com uma equipe que abarca quatro universidades, foi organizado de forma a obter amostras das seguintes áreas: Rio Grande do Sul: Porto Alegre, Flores da Cunha, Panambi e São Borja; Santa Catarina: Florianópolis, Blumenau, Chapecó e Lages; Paraná: Curitiba, Londrina, Irati e Pato Branco.

Os dados são estratificados e obtidos seguindo-se o modelo laboviano de condução de entrevistas. Sua transcrição dá-se em três linhas: transcrição ortográfica na primeira linha, indicação de variação na segunda linha com a finalidade de se ligar eletronicamente a forma ortográfica da primeira linha com suas realizações, e, na terceira linha, classificações morfossintáticas e registros de estilo de fala (Bisol, 2005; Costa, 2005).

O projeto VARSUL atualmente é integrado por três subprojetos: Banco de Dados VARSUL, Amostra Digital VARSUL e Banco de Dados Diacrônico. Ao longo dos anos, o projeto produziu muitas publicações, teses e dissertações; muitos desses trabalhos são acessíveis através do *site* do projeto. O VARSUL continua ampliando seu espectro de variedades e atualmente disponibiliza em seu *site* uma amostra digital,[49] cujo projeto é também de livre acesso.[50] A amostra digital é composta por gravações em formato wav e suas transcrições; adicionalmente é oferecida a possibilidade de ligação entre o som e o texto transcrito via *software* Oceanaudio.

2.8 ALIP: Projeto Amostra Linguística do Interior Paulista – o banco de dados IBORUNA

O banco de dados IBORUNA,[51] compilado sob a coordenação de Sebastião Carlos Leite Gonçalves, tem como objetivo fornecer dados para a análise do português falado na região noroeste do interior paulista, cobrindo os seguintes municípios: Bady Bassit (BAD), Cedral (CED), Guapiaçu (GUA), Ipiguá (IPI), Mirassol (MIR), Onda Verde (OND) e São José do Rio Preto (SJP).

---

[49] <http://www.varsul.org.br/?modulo=pagina&id=47>.

[50] <http://chu.com.br/varsul/downloads/projeto_amostra_digital_varsul.pdf>.

[51] <http://www.iboruna.ibilce.unesp.br/>.

O *corpus* constitui-se de dois tipos de amostras: Amostra Comunidade (ou Amostra Censo - AC) e Amostra de Interação Dialógica (AI). A AC é composta por 152 amostras controladas sociolinguisticamente. A cada amostra correspondem cinco arquivos sonoros, um arquivo com dados da ficha social do informante, um arquivo de transcrição e um arquivo com registros do diário de campo. A AI agrupa amostras de fala secretamente gravadas em situações de interação social. A cada amostra corresponde um arquivo sonoro, um arquivo com dados da ficha social dos informantes, um arquivo de transcrição e um arquivo com registros do diário de campo.

O banco de dados foi financiado integralmente pela FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), que, além dos equipamentos necessários para a execução do projeto, subsidiou também bolsas de treinamento técnico para os membros da equipe técnica, responsáveis pela coleta das amostras de fala.

Gonçalves (2008) apresenta o projeto detalhadamente, discutindo os parâmetros para a sua compilação, processos de validação e transcrição, além das dificuldades enfrentadas em diferentes etapas do projeto, bem como uma avaliação dos resultados obtidos. O banco de dados IBORUNA é de acesso livre, disponibilizado através do *site* do projeto.

3. O C-ORAL-BRASIL

O C-ORAL-BRASIL[52] é um *corpus* de fala espontânea do português brasileiro, representado pela diatopia mineira, majoritariamente proveniente da região metropolitana de Belo Horizonte. Os parâmetros de compilação desse *corpus* são pautados pela arquitetura proposta no C-ORAL-ROM (Cresti; Moneglia, 2005), um conjunto de *corpora* das quatro principais línguas românicas europeias: espanhol, francês, italiano e português europeu.

O projeto, iniciado em 2007 com um estudo piloto, apresenta neste volume o produto de sua primeira fase de compilação, isto é, a parte informal do C-ORAL-BRASIL. O detalhamento técnico do *corpus*, assim como a sua descrição, são minuciosamente discutidos nos capítulos 2 e 4 deste volume.

Os objetivos da compilação de um *corpus* dessa natureza são múltiplos: estudos da estrutura da fala espontânea, estudos da organização informacional da fala, estudos de fenômenos segmentais e suprassegmentais da fala, estudos de caracterização de variação linguística, estudos morfossintáticos,

[52] <http://www.c-oral-brasil.org>.

estudos semânticos, dentre outros. A fala espontânea aqui é caracterizada por Moneglia (2005, p. 4), através dos seguintes atributos: ocorre em interações multimodais face a face, apresenta referência intersubjetiva a um espaço dêitico, programação simultânea à execução e comportamento linguístico contextualmente indeterminado (comportamento imprevisível).

O C-ORAL-BRASIL é um *corpus* balanceado, representativo das três principais tipologias da fala – diálogos, monólogos e conversações –, em contextos públicos e privados. Sua amostragem é de cerca de 1.500 palavras por texto. As especificações do *corpus* seguem as orientações internacionais do Text Encoding Initiative (TEI)[53] e usufruem das especificações técnicas previamente testadas e recomendadas pelo C-ORAL-ROM para compilação de um *corpus* representativo da fala espontânea (cf. Cresti; Moneglia, 2005). Cada um dos textos gravados é apresentado no *corpus* através de seu arquivo de som, em formato wav, acompanhado de sua transcrição em formato rtf, devidamente segmentada em unidades tonais (quebras terminais e não terminais),[54] e de seu arquivo de alinhamento texto-som em formato xml, obtido via *software* WinPitch (Martin, 2005). Cada um dos textos apresenta também o seu arquivo de metadados, seguindo os padrões TEI, e suas versões txt e xml etiquetadas morfossintaticamente via anotador sintático PALAVRAS (Bick, 2000).

A documentação da fala espontânea suscita uma série de dificuldades que se iniciam, como anteriormente indicado, na gravação dos informantes. Todas as gravações efetuadas no âmbito do C-ORAL-BRASIL foram feitas com a ciência dos informantes, que assinaram o termo de consentimento informado, aprovado pelo Comitê de Ética da UFMG. As situações de gravação foram as mais variadas possíveis, incluindo desde a execução de tarefas domésticas a trabalho em obra de construção civil, partida de futebol, garçons em uma festa, dentre outras. Não houve nenhuma gravação de fala roteirizada ou direcionada por tarefas propostas pela equipe do projeto. Essa multitude de cenários interativos garantiu a representatividade diafásica do *corpus*. A possibilidade de gravações tão variadas só foi possível devido à alta qualidade do equipamento utilizado pela equipe, que inclui microfones sem fio e *mixer*. Apenas as gravações classificadas como sendo de alta qualidade acústica foram utilizadas. Para se maximizar a qualidade

---

[53] <http://www.tei-c.org/index.xml>.

[54] Cf. capítulos 2 e 4 para uma discussão detalhada da segmentação da cadeia sonora em unidades de referência da fala.

das gravações, toda a equipe foi treinada através de oficinas para a adequada utilização dos equipamentos.

O segundo passo no processo de compilação do *corpus* foi o processo de transcrição dos arquivos de som. Novamente, a equipe foi extensivamente treinada para efetuar a transcrição e segmentação em unidades tonais do sinal sonoro. A transcrição segue critérios estabelecidos em um protocolo pelos coordenadores do projeto, levando-se em conta fenômenos de lexicalização e gramaticalização em curso no português brasileiro. As diretrizes de transcrição procuram se ater maximamente aos padrões ortográficos da língua portuguesa, apenas adicionando formas necessárias seja pela sua não dicionarização ou pela sua significância sistêmica nos níveis fonológico ou morfossintático (cf. Mello; Raso, 2009), atentando-se aos parâmetros propostos pelo Grupo de Trabalho de Língua Oral – EAGLES.[55] A codificação das transcrições seguiu o padrão CHAT (MacWhinney, 2000).

O trabalho de etiquetagem morfossintática do *corpus* foi desenvolvido por Eckard Bick através de adequações especialmente desenhadas para o tratamento da fala espontânea do português brasileiro e implementadas no processador morfossintático PALAVRAS. O detalhamento desse trabalho é relatado por Bick no Capítulo 6 deste volume, em que são discutidas as decisões relacionadas às estruturas formais para a análise sintática (unidades tonais) e codificação de formas lexicais presentes somente na fala.

Além dos aspectos inovadores relacionados à documentação e tratamento da fala espontânea do português brasileiro, sua representatividade e balanceamento, o C-ORAL-BRASIL é um *corpus* validado em todos os seus estágios de compilação. Os aspectos relacionados à validação da transcrição e segmentação prosódica são discutidos no Capítulo 4 deste volume. Já aqueles relacionados à etiquetagem morfossintática são apresentados no Capítulo 6. Os resultados das validações estatísticas são muito positivos e atribuem um alto grau de confiança e acuidade a esse *corpus*, estabelecendo-o, dessa forma, como um recurso de referência a ser utilizado pela comunidade científica. Tal grau de acuidade foi alcançado devido ao conjunto de passos metodológicos adotados no seu processo de compilação, pautados na experiência inovadora prévia da equipe executora do C-ORAL-ROM (Cresti; Moneglia, 2005) e na adequação e ajuste de seus parâmetros para a realidade do português brasileiro. Ademais,

[55] <http://www.ilc.cnr.it/EAGLES/home.html>.

o esforço metodológico para se atingir um alto grau de confiabilidade para o C-ORAL-BRASIL contou com várias etapas de revisão de todo o trabalho executado, em momentos sucessivos, até serem alcançados os patamares de excelência estabelecidos pelo projeto. O próximo passo nesse processo é a validação externa ao projeto, a ser feita por uma equipe de *experts*. Tal validação é protocolo usual dos grandes projetos de *corpora* internacionais e visa garantir a sua confiabilidade através da avaliação do nível de correspondência entre a sua descrição e o produto final alcançado.

4. Considerações finais

A linguística brasileira tem-se preocupado com os estudos da língua em uso há várias décadas, como documentado pelos esforços de pesquisa e representação da fala já protagonizados na década de 1960, quando as tendências teóricas predominantes ainda favoreciam as metodologias introspectivas como principal forma de acesso aos dados linguísticos. A tradição de estudos dialetológicos, que esteve fortemente presente nos primórdios da linguística brasileira, e o seu enriquecimento com as metodologias de estudo da Sociolinguística Variacionista, garantiram o crescente interesse pelo estudo de dados linguísticos reais no Brasil. Tal vocação foi marcada pelos inúmeros trabalhos desenvolvidos e orientados por Nelson Rossi e por seus ex-alunos, assim como por aqueles desenvolvidos no âmbito do projeto NURC em todo o país, e também pelo polo irradiador de formação em Sociolinguística Variacionista e estudos funcionalistas da linguagem representado pela equipe da UFRJ.

Nesse sentido, da década de 1960 até o ano de 2000, firmou-se no país uma sólida tradição de estudos baseados no uso linguístico. O século XXI nos trouxe novas tecnologias e metodologias para o estudo da língua, assim como muito mais disponibilidade de fomento à pesquisa e sensibilidade à temática da compilação de *corpora* por parte das agências financiadoras no Brasil. Agora, ancorada pelas propostas que já vêm se desenvolvendo há décadas no exterior, a linguística brasileira tem expandido as suas possibilidades de pesquisa ao adotar as metodologias da Linguística de *Corpus*, campo disciplinar que agrega inovações paradigmáticas calcadas na interlocução dos estudos linguísticos com princípios da Estatística e da Ciência da Computação.

A organização de dados linguísticos em *corpora*, nesse campo disciplinar, não tem caráter apenas de formação de banco de dados linguísticos.

Nesse novo momento da disciplina Linguística, os dados devem ser agrupados de acordo com desenhos metodológicos estabelecidos previamente à coleta, visando à sua representatividade, de forma balanceada e com amostragens adequadas, a fim de que haja uma margem estatística segura de acerto a qualquer estudo neles calcados. Já não há mais a possibilidade de se considerarem plausíveis técnicas de coleta de dados ou de processamento que deixem margem à sua contaminação ou criação de tendências exógenas à língua.

O rigor metodológico é o maior trunfo da Linguística de *Corpus*, e sem ele não se podem compilar *corpora*. Assim, na compilação de *corpora* de fala espontânea, há que se observarem os parâmetros contingentes ao tratamento de dados dessa diamesia, a fim de que se obtenham *corpora* representativos da estrutura e particularidades da língua falada que a difiram muito significativamente da língua escrita.

É nesse cenário, de proposição metodológica crucialmente refinada, que o C-ORAL-BRASIL foi pensado e compilado. O *corpus*, produto de pesquisa intensa e grande esforço metodológico, é inovador no panorama da ciência linguística brasileira, e o que inaugura, esperamos, será um passo adiante na descrição e compreensão do português brasileiro.