NURC Digital

Um protocolo para a digitalização, anotação, arquivamento e disseminação do material do Projeto da Norma Urbana Linguística Culta (NURC)

Miguel Oliveira, Jr.
Universidade Federal de Alagoas, Maceió, Brazil

This paper presents the procedures that have been adopted for the digitization, annotation, archiving and dissemination of a corpus of the NURC - Norma Linguística Urbana Culta [*Cultured Linguistic Urban Norm*] Project, an academic project that assembled one of the most influential speech corpora for linguistics research in Brazil. The goal here is to make public a protocol that was created based on the above-mentioned procedures, and to promote it as best practices to be adopted for digitizing linguistic data that were recorded in analog format, in general, and, in particular, for the digitization of the NURC Project data; and for the annotation, archiving and dissemination of linguistic data in digital format.

Keywords: NURC, digitalization, annotation, archiving, dissemination

1. Introdução

O Projeto da Norma Urbana Linguística Culta teve seu início em 1969, tendo sido proposto como uma extensão do *Proyecto de Estudio Coordinado de la Norma Linguística Culta de las Principales Cidades de Iberoamérica y de la Península Ibérica*, de que participavam países de língua espanhola da América Latina. A proposta inicial do Projeto era documentar e estudar a norma falada culta de cinco capitais brasileiras: Recife, Salvador, Rio de Janeiro, São Paulo e Porto Alegre. A seleção dessas capitais foi feita a partir dos seguintes critérios: ter a cidade pelo menos um milhão de habitantes e estratificação social suficiente para atender às exigências do projeto.

CHIMERA. Romance Corpora and Linguistic Studies 3.2 (2016), 149-174. ISSN 2386-2629

Os dados que fazem parte do acervo do Projeto NURC têm sido utilizados para a elaboração de um grande número de trabalhos acadêmicos, incluindo dissertações de mestrado, teses de doutorado, artigos publicados em periódicos nacionais e internacionais, e trabalhos apresentados em encontros científicos ao redor do mundo. A Gramática do Português Falado (Castilho 1990; Castilho 1993; Castilho & Basílio 1996; Ilari 1992; Kato 1996; Koch 1996; de Moura Neves 1999; Abaurre & Rodrigues 2002), grande e ambicioso projeto nacional que envolveu entre 1988 e 2002 cerca de cinquenta pesquisadores na área da linguística, resultou em uma série de volumes, todos contendo análises de materiais extraídos dos dados do Projeto NURC. É, pois, incontestável a importância do material pertencente ao arquivo do Projeto NURC.

Lamentavelmente, os registros magnéticos dos inquéritos do Projeto NURC, feitos em fita de rolo, estão em sério risco de deterioração. Como se sabe, fitas magnéticas estão sujeitas a uma série de degradação, sendo as principais delas a hidrólise e oxidação, o que compromete a sua vida útil (Van Bogart 1995). Os mais representativos centros de pesquisa e documentação ao redor do mundo possuem registros em formato de áudio. Estes materiais, mantidos na forma analógica, tem uma expectativa de vida reduzida em relação aos materiais registrados em suportes de papel. Quanto maior a demora em se migrar estes dados para meios digitais, maior é o risco de perda dos conjuntos de som, especialmente em países de clima tropical como o Brasil. Muitos dos registros do Projeto NURC já se encontram irremediavelmente destruídos pela ação do tempo. É imprescindível, portanto, que este valioso material seja resgatado o quanto antes, mediante a transposição de seus dados analógicos para formatos digitais que garantam a sua preservação e utilização no futuro.

O objetivo central do projeto do Projeto NURC Digital, financiado pelo CNPq, Chamada Universal 14/2012, Processo no 472918/2012-5, foi desenvolver uma metodologia e práticas específicas para gestão dos registros sonoros resultantes das atividades do Projeto NURC, bem como de estratégias de migração para formatos digitais, curadoria e preservação digital do acervo. A pesquisa visou estudar meios que poderão ser utilizados pelo Projeto NURC em todas as capitais em que está sediado para a preservação e a disponibilização mais efetiva de seus corpora. Para isso, o projeto gerou um protocolo de digitalização, anotação, arquivamento e disseminação do material do Projeto NURC, feito a partir da adoção de um corpus representativo de inquéritos pertencentes ao acervo do Recife. Este protocolo levou em consideração técnicas de digitalização e de arquivamento recomendadas por órgãos internacionais especializados em arquivamento de dados digitais.

Este artigo apresenta os procedimentos que foram adotados para a proposição do protocolo de digitalização, anotação, arquivamento e disseminação do material do Projeto NURC e descreve o protocolo em si. O objetivo aqui é tornar público o protocolo do Projeto NURC Digital, com o propósito de promovê-lo entre a comunidade como boas práticas a serem adotadas na digitalização de dados analógicos em geral, e dos dados do Projeto NURC em particular.

2. Estado da arte

Entende-se por corpus, nos estudos linguísticos, uma "coletânea de porções de linguagem que são selecionadas e organizadas de acordo com critérios linguísticos explícitos, a fim de serem usadas como uma amostra da linguagem" (Percy *et al.* 1996: 4). O corpus do Projeto NURC é uma coletânea de dados de fala de informantes com formação universitária completa (chamados cultos), organizada para servir de estudo da modalidade oral da língua portuguesa culta falada no Brasil.

O projeto original previa três etapas de trabalho: (i) gravações de dados (ou inquéritos, como passaram a ser chamadas as sessões de gravação de dados do Projeto), (ii) transcrição dos dados gravados, utilizando-se para isso normas de transcrição derivadas da Análise da Conversação, e (iii) estudos linguísticos baseados no corpus, abrangendo diferentes níveis de análise.

Inicialmente, foram previstas 400 horas de gravação, selecionando-se 600 informantes (300 do sexo masculino e 300 do sexo feminino) com nível superior de escolaridade, nascidos na cidade sob estudo ou nela residentes desde os cinco anos de idade, filhos de nativos de língua portuguesa, de preferência nascidos na cidade sob pesquisa. Os informantes foram distribuídos em três faixas etárias: (i) 1ª faixa etária: de 25 a 35 anos de idade; (ii) 2ª faixa etária: de 36 a 55 anos de idade; (iii) 3ª faixa etária: acima de 56 anos de idade. Quanto à natureza, as gravações foram divididas em quatro tipos: (i) Gravações secretas de um diálogo espontâneo (GS); (ii) Diálogo entre dois informantes (D2); (iii) Diálogo entre o informante e o documentador (DID); (iv) Elocuções Formais (EF). As gravações secretas nunca chegaram a ser realizadas devido à conjuntura política em que o Brasil se encontrava na época de implementação do corpus.

O material do Projeto NURC foi – e tem sido – largamente utilizado para o estudo de diversas características da oralidade, que vão desde aspectos discursivos, tais como a análise de narrativas inseridas na conversação (Oliveira 1999) e de questões discursivas e ideológicas presentes nas diversas modalidades de gravações feitas pelo Projeto (da Cunha 2003), até aspectos

mais formais, tais como a análise de elementos argumentativos e pragmáticos, da intertextualidade e da organização interacional e sintática presentes no texto oral (Sá 2004).

A maior parte dos estudos desenvolvidos a partir dos dados do Projeto NURC deriva de uma série de publicações feitas com transcrições de material selecionado pelos grupos de pesquisadores atuantes em cada uma das capitais em que o Projeto era desenvolvido. Essas coletâneas de transcrições publicadas a partir da década de 80 ficaram conhecidas por *Materiais Para o Seu Estudo* (Castilho 2007): Castilho & Preti (1986, 1987), Preti & Urbano (1990), Callou (1992), Callou & Lopes (1993, 1994), Motta & Rollemberg (1994), Hilgert (1997), Sá *et al.* (1996, 2005). Os estudos feitos a partir dessas publicações desconsideravam, em sua grande maioria, o registro de áudio, baseando-se exclusivamente nas transcrições aí presentes. Essa não era, evidentemente, uma opção dos estudiosos. Tratava-se mesmo de uma questão de dificuldade de acesso aos dados gravados. Todas as gravações feitas pelo Projeto NURC utilizaram, como meio, fitas magnéticas de rolo, que, se por um lado garantia a qualidade das gravações, por outro dificultava o acesso às mesmas, uma vez que reprodutores de fita de rolo eram equipamentos caros e pouco comuns.

Uma outra dificuldade que a utilização do material do Projeto NURC apresentava aos estudiosos era – e continua sendo, em grande parte – a não disponibilização dos dados transcritos em formato digital. Assim, o processo de análise a partir dos textos publicados em formato impresso era – e continua sendo – necessariamente demorado e eventualmente falho, uma vez que não se podia contar com buscas automatizadas de fenômenos linguísticos particulares.

Com o advento da tecnologia, tem-se cada vez mais incentivado a disponibilização de dados linguísticos em formato digital, que possam ser acessados por humanos e máquinas. A simples digitação de dados é apenas um primeiro passo para a criação de um corpus digital. Há, na verdade, uma série de medidas recomendadas por especialistas na área da construção de corpora eletrônicos que precisam ser consideradas, se o objetivo for construir um corpus que seja também legível por máquinas (Sardinha 2000)[1]. A vantagem de se construir um corpus com essa característica é mesmo a de facilitar as análises linguísticas feitas a partir dele, automatizando certos aspectos da análise. À análise linguística que toma por base corpora informatizados para deles fazer considerações probabilísticas tem-se comumente referido como linguística de corpus (Sardinha 2000).

[1] Muitas dessas medidas serão apresentadas e discutidas mais adiante.

Já houve tentativas isoladas de informatização de dados do Projeto NURC (Castilho *et al.* 1995). Assim, por exemplo, muitos dos dados do Projeto NURC do Rio de Janeiro foram digitalizados e disponibilizados na internet[2]. A despeito de ser essa uma empreitada louvável, a metodologia empregada para a disponibilização desses dados online não levou em consideração uma série de recomendações metodológicas bastante importantes no processo de elaboração de bancos de dados digitais. Desse modo, apesar de agora pesquisadores interessados em aspectos da oralidade poderem ter acesso aos arquivos de áudio a que se referem algumas transcrições, e poderem fazer buscas bastante rudimentares no corpus disponibilizado pelo NURC-RJ, não poderão, entre outras coisas, proceder, por exemplo, a uma análise automatizada de frequência de ocorrência de traços linguísticos de várias ordens (lexicais, sintáticos, semânticos, discursivos, etc.), ou a uma possível análise acústica, devido à não-observação das já referidas recomendações metodológicas.

A área da linguística que tem se preocupado em estabelecer bases teóricas para a construção de corpora linguísticos digitais é chamada linguística documentativa (Himmelmann 2006). A linguística documentativa emergiu como uma resposta para uma necessidade urgente de se fazer registros duradouros de línguas em risco de extinção, utilizando-se o aparato tecnológico disponível na atualidade. Entretanto, a sua área de atuação hoje em dia vai além da documentação de línguas em risco de extinção. A linguística documentativa se ocupa em indicar métodos e ferramentas para a elaboração de registros de qualquer língua natural, ou de variedades de uma língua, que sejam representativos, duradouros e que permitam múltiplos usos.

Os procedimentos estabelecidos para a construção de um corpus linguístico digital permitem a sua utilização não apenas em diversas áreas da linguística, tais como a fonologia, a fonética, a morfologia, a sintaxe, a semântica, a análise do texto e do discurso, a sociolinguística, a tipologia, etc., mas também em áreas afins, como a história (história oral), a antropologia (aspectos culturais, questões acerca da interação), a sociologia, a poética (aspectos musicais e métricos da literatura oral), e a educação (estudo de gêneros da oralidade em sala de aula), por exemplo.

Outro aspecto importante do Projeto NURC Digital é sua inserção nas áreas da engenharia e computação. Os resultados alcançados serão de valia para a construção de sistemas computacionais que realizem o processamento automático da fala. Por exemplo, tanto o reconhecimento quanto a síntese de voz são importantes tecnologias assistivas, que podem melhorar substancialmente a

[2] http://www.letras.ufrj.br/nurc-rj/home.htm

qualidade de vida de pessoas com necessidades especiais, tais como as com deficiência auditiva ou visual. Tais tecnologias são *data-driven* e dependem de grandes bases de dados, devidamente rotuladas, para o adequado desenvolvimento de sistemas. Além do aspecto econômico, o processamento de voz é um dos melhores exemplos de um relevante desafio enfrentado pela computação (e áreas afins): a construção de máquinas capazes de interagir de forma natural com seres humanos.

Apesar da reconhecida importância, as atividades em processamento de fala no Brasil, tanto na academia quanto na indústria (em especial na indústria de *software*) ainda não alcançam a dimensão necessária para que as mesmas tragam benefícios significativos à sociedade. Isso acontece porque tanto reconhecimento quanto síntese de voz são tecnologias que dependem de grandes bases de dados, devidamente rotuladas, para o adequado desenvolvimento de sistemas no estado-da-arte. Em outras palavras, para atingir altas taxas de acerto é imprescindível uma grande base de voz digitalizada, rotulada e que tenha cobertura de todas as variações na fala no Brasil. Contudo, há poucas ações nesse sentido. O Projeto NURC Digital foi proposto com o objetivo de colabora para diminuir essa lacuna.

Além de propor uma metodologia de organização de um corpus representativo do acervo do Projeto NURC, em formato digital, que poderá servir como modelo a ser adotado para a informatização de todo o material pertencente ao arquivo do Projeto NURC, os resultados do projeto NURC Digital beneficia diretamente a comunidade científica, que passará a ter disponíveis para consulta otimizada dados – anteriormente de difícil acesso – em formato digital de alta qualidade, devidamente catalogados, etiquetados e transcritos. Alguns dos objetivos específicos do projeto foram: (i) digitalizar todo o acervo do Projeto NURC/Recife, incluindo não apenas os arquivos de áudio, mas também as informações referentes ao material de áudio digitalizado (metadados) e os dados de transcrição feitos segundo as normas do Projeto NURC; (ii) propor um sistema de anotação/etiquetagem multi-nível para os dados do Projeto NURC; (iii) anotar/etiquetar um corpus representativo dos dados do Projeto NURC (o corpus compartilhado do Projeto NURC/Recife), tornando-os alinhados[3], o que propiciará uma utilização mais proveitosa dos mesmos, com informações multi-níveis; (iv) disponibilizar para a comunidade acadêmica, em site dedicado, todo o material do Projeto NURC/Recife, incluindo o corpus compartilhado amplamente anotado, que poderá ser acessado

[3] O conceito de anotação/etiquetação alinhada será discutido mais adiante.

mediante um sistema de busca avançada e (v) arquivar os dados informatizados em bancos de dados internacionais, assegurando assim a sua preservação.

Na seção abaixo, serão discutidas questões metodológicas que foram consideradas para a elaboração do protocolo que aqui se apresenta.

3. Considerações metodológicas

Os inquéritos do Projeto NURC foram gravados em condições variadas. Em geral, as gravações eram realizadas com microfones dinâmicos omnidirecionais, apoiados em uma mesa. Todos os inquéritos foram registrados em fitas magnéticas de rolo profissionais, com espessura de 0,0018mm, largura de 6,35mm e comprimento de 540m (BASF TP 18 LH). Os equipamentos utilizados para a gravação foram os seguintes gravadores de rolo: AKAI 4000 DS Mk –II, SONY TC – 366 e Philips N 4416, sendo o primeiro deles o mais utilizado. A depender do tipo de inquérito, as gravações eram realizadas em salas específicas, em salas de aula, em auditórios e, em alguns casos, nas casas dos próprios informantes. Portanto, a qualidade acústica das gravações do Projeto NURC é bastante heterogênea.

Para a digitalização dos dados do Projeto NURC/Recife, foram observadas as recomendações propostas pelo *Open Archival Information System* (OAIS), que é um modelo de referência, com padrão ISO (14721:2003), adotado pelos bancos digitais de dados linguísticos mais recentes, e pelo Comitê Técnico da *International Association of Sound and Audiovisual Archives* para objetos digitais (Bradley 2009; Von Arb & Gaustad 2005).

A digitalização é um processo de conversão de um sinal elétrico (analógico, contínuo), em informações expressas em números (discretas, descontínuas). O conversor AD (analógico-digital) transforma os sinais elétricos em números por um processo de amostragem. Neste processo, as amostras são medidas em intervalos fixos. O número de vezes em que se realiza a amostragem em uma unidade de tempo é a *taxa de amostragem*, geralmente medida em Hertz. De acordo com o Teorema de Nyquist, a taxa de amostragem deve ser maior que o dobro da maior frequência contida no sinal a ser amostrado, para que ele possa ser reproduzido integralmente. Como ouvimos numa faixa que vai aproximadamente de 20 a 20kHz, a taxa de amostragem deveria ser, de acordo com o que propõe o Teorema de Nyquist, maior que o dobro da maior frequência desse espectro, para que ele possa ser reproduzido integralmente. Assim, em teoria, e levando-se em conta a percepção auditiva humana, os sons devem ser amostrados numa frequência superior a 40kHz para que todas as

frequências audíveis possam ser registradas. Parece, no entanto, haver um consenso geral de que, embora os seres humanos não ouçam componentes de frequência acima de 20 kHz, o sinal analógico tende a ser representado de forma mais precisa e com menos ruído em taxas de amostragem maiores, como a de 96.000 Hz, que é hoje em dia o padrão da indústria de cinema e áudio profissional. Segundo Plichta & Kornbluh (2002) utilizar taxas de amostragem mais elevadas resulta em um aumento da frequência de resposta e melhora a relação sinal ruído.

Um outro conceito muito importante para o processo de digitalização é o de *resolução*. A quantidade de valores possíveis para indicar a amplitude em cada amostra do sinal elétrico/acústico a ser convertido é expressa, ou codificada, em *bits*. O computador, para fazer seus cálculos, usa o sistema binário: um bit é um instante em que ele verifica a presença ou ausência de eletricidade; cada um desses casos é representado com os dígitos zero e um. A resolução define quantos bits são usados para descrever cada uma das amostras colhidas a partir da taxa de amostragem. Quanto maior for a resolução do sinal de áudio, mais detalhes acerca das rápidas mudanças de amplitude do som terá o sinal digital resultante. As taxas de resolução mais comuns são as de 8, 16 e 24 bits. Resoluções de 8 bits ocupam menos espaço mas produzem arquivos de som com mais ruídos. Uma taxa de resolução de 24 bits, ajuda a capturar mais detalhes e a minimizar o ruído de digitalização e a distorção (Plichta & Kornbluh 2002).

A *International Federation of Library Associations and Institutions* (IFLA) e a *International Association of Sound and Audiovisual Archives* (IASA) recomendam que se busque oferecer a melhor qualidade possível para arquivo digital quando se planeja disponibilizar ao acesso largo materiais de áudio através da digitalização. Levando em consideração essas questões, e o fato de que a tecnologia atual permite a utilização de taxas de amostragem mais altas e uma maior gama de resolução em número de bits – algo proibitivo há muito pouco tempo, por questões relativas a espaço de armazenamento, elegemos como padrão de digitalização dos dados do Projeto NURC taxa de amostragem de 96.000 Hz e quantização de 24 bits.

Uma série de testes com diferentes equipamentos foi realizada para encontrar a melhor solução de hardware para a digitalização dos dados de áudio. A seleção de equipamentos a serem testados baseou-se em especificações técnicas e critérios discutidos pela literatura (Burg *et al.* 2014). Como os arquivos de áudio do Projeto NURC são relativamente longos (em média, cada gravação dura 40min), especificações técnicas do computador, como o processador, a memória embarcada e a capacidade de armazenamento foram consideradas essenciais. No que diz respeito aos outros itens de hardware, as

seguintes características foram levadas em consideração: (i) especificações técnicas adequadas, (ii) intercompatibilidade e (iii) robustez.

Também uma bateria de testes foi realizada para encontrar software que melhor se adequasse às necessidades do projeto. Na seleção de programas a serem testados, foi dada preferência àqueles gratuitos e de linguagem aberta. Os principais critérios utilizados nos testes de software foram: (i) a interface do aplicativo e a facilidade de uso, (ii) os recursos de gravação e de edição, (iii) os recursos de restauração e (iv) as especificações de exportação (formato e qualidade do arquivo de áudio).

Seguindo recomendações dos órgãos supracitados, todos os arquivos de áudio foram salvos em formato não-comprimido, baseado em PCM (*Pulse Code Modulation*, ou Modulação por Código de Pulsos), de maneira a garantir a qualidade os dados digitalizados, preservando o maior número de informações possíveis presentes no sinal analógico.

As informações referentes aos arquivos de texto, às transcrições e aos metadados foram digitalizadas seguindo as recomendações do OAIS e da IFLA. Em particular, os metadados foram organizados a partir dos padrões propostos pela *ISLE Meta Data Initiative* (IMDI) e pelo *Text Encoding Initiative* (TEI), adotados por vários bancos de dados internacionais. As transcrições, por sua vez, foram encapsuladas em formatos abertos e transparentes, que podem ser gerados e lidos em virtualmente qualquer plataforma operacional e suportam o unicode (padrão que permite a representação e manipulação de texto de qualquer sistema de escrita).

A transcrição de dados de fala é uma atividade de extrema importância e que exige bastante cuidado. Em vários corpora linguísticos, a transcrição é ainda o único produto que chega a ser acessível. Portanto, é fundamental que seja cuidadosamente pensada, trabalhada e revisada. É um fato reconhecido que a qualquer transcrição linguística subjaz uma teoria. A escolha da unidade de análise (o segmento, a palavra, a unidade entoacional, a frase, o turno), por exemplo, já pressupõe uma decisão teórica. Portanto, a postura teórica que se adota na transcrição da fala é fundamental para garantir consistência ao corpus (Edwards & Lampert 1992). Antes de dar início a qualquer processo de transcrição, é necessário delimitar o grau de detalhamento da transcrição e decidir que aspectos serão sempre e sistematicamente registrados, e como esses aspectos serão registrados. Ramilo & Freitas (2010: 68) afirmam que "todas as decisões tomadas relativamente ao método de transcrição influem no resultado final do projeto e no seu posterior aproveitamento". O objetivo de uma transcrição linguística é transportar o discurso falado para registros gráficos da forma mais fiel possível. Entretanto, qualquer transcrição é necessariamente

descontínua, porque precisa recorrer a elementos discretos para representar o que se manifesta continuamente e dissociativa, porque não consegue reproduzir todos os componentes associados ao discurso falado (Paiva 2003). Embora hoje em dia já estejam popularizados muitos tipos de anotação, os corpora orais têm historicamente privilegiado a transcrição ortográfica, porque este tipo de transcrição facilita a legibilidade de humanos e máquinas.

As normas de transcrição do Projeto NURC, que podem ser conferidas nos volumes *Materiais para o seu estudo*, são adotadas em vários corpora, sendo um modelo bastante utilizado e bastante influente no Brasil. O modelo, proposto por Marcuschi (1986), é baseado no sistema de transcrição desenvolvido por Sacks *et al.* (1978) para a Análise da Conversação, que foi ampliado, mais tarde, por MacWhinney (2000). Esse modelo privilegia a transcrição ortográfica, mas apresenta inconsistências. A principal delas diz respeito à tentativa de abarcar fenômenos suprassegmentais, tais como a pausa, a entoação, o alongamento de vogais e a ênfase, sem adotar, algumas vezes, parâmetros claros para isso.

É importante, entretanto, salientar que a transcrição ortográfica é apenas uma anotação possível de dados de fala. Como mencionado acima, um dos objetivos do projeto NURC Digital é propor um sistema de anotação multinível para os seus dados. Anotar um texto significa adicionar informações a suas partes constitutivas a partir de uma análise das mesmas. Um tipo comum de anotação é a indicação de categorização, por meio de etiquetas. Assim, por exemplo, a categorização de palavras em termos de classes gramaticais/morfológicas. Sabe-se que um corpus anotado facilita enormemente o trabalho de análise, a despeito de possíveis problemas que nele existam.

Há diferentes tipos de anotação: fonética, fonológica, prosódica, morfológica, sintática, semântica, pragmática, discursiva, etc. Em princípio, quanto mais anotado um corpus em diferentes níveis, maior será a sua aplicabilidade. Entretanto, é importante enfatizar que um corpus anotado só será útil se amparado em uma anotação bem planejada e cuidadosamente executada. Uma das principais recomendações metodológicas feitas para anotação de corpora é que toda anotação deve vir em separado, como um componente extra ao corpus. É imprescindível que um usuário possa ter acesso ao corpus não-tratado, em seu estado bruto. Informações explícitas e detalhadas sobre o sistema de anotação adotado devem ser disponibilizadas.

Para a seleção do sistema de anotação a ser utilizado, é preciso que se considere um sistema que seja consensual, para que a anotação seja utilizada/compreendida por um maior número de usuários. Há muitos sistemas de anotação disponíveis; deve-se optar por aquele que tem mais tradição na comunidade científica, para que os dados anotados possam ser comparáveis.

Como bem apontam Ramilo & Freitas (2010), uma das características compartilhadas por boa parte dos corpora de fala organizados nos últimos anos é o fato de apresentarem alinhamento das anotações com o áudio. Uma das maiores vantagens deste tipo de anotação é o fato de agilizarem o processo de localização de excertos áudio na utilização final do corpus. Existem diferentes formas de produzir o alinhamento manual das anotações. Hoje é possível contar com um número razoável de aplicativos computacionais que auxiliam o processo de anotação alinhada, entre os quais podemos citar o *Praat*[4], o *ELAN* (Wittenburg *et al.* 2006), o *WinPitch* (Martin 2004) e o *TranscriberAG*[5]. Alguns desses programas oferecem a possibilidade de uma anotação assistida; outros exigem que o alinhamento da anotação seja feito no próprio programa.

Qualquer que seja o método utilizado para a produção de uma anotação alinhada, é preciso que se estabeleça o critério de segmentação a ser adotado. A tarefa de anotação alinhada pode ser executada com maior ou menor precisão, e ter como alvo diferentes unidades linguísticas. Em geral, a depender do tamanho do corpus a ser anotado, a opção recai por unidades maiores que a palavra, uma vez que quanto menor a unidade linguística adotada para a segmentação, maior o dispêndio de tempo para a execução da anotação. A eleição da unidade de segmentação a ser adotada, como já mencionado acima, vai depender da orientação teórica a que os executores do projeto afiliam-se. Tem sido uma prática corrente, todavia, em corpora orais, a adoção de unidade de segmentação que leva em conta aspectos prosódicos do enunciado, como a unidade entoacional.

A unidade entoacional indica a maneira como um dado falante emoldura o seu enunciado. Não há consenso na literatura acerca da definição de unidade entoacional. As definições vão desde considerações fisiológicas (Lieberman 1967) e cognitivas (Chafe 1994) até abordagens semânticas (Halliday 2004) e gramaticais (Selkirk 1984), encontrando todas essas sérias limitações. De acordo com Crystal (1969), considerações fonológicas são em geral suficientes para definir o que chama unidade tonal. Cruttenden (1986), por sua vez, propõe que informações fonéticas, fonológicas, semânticas e gramaticais devem ser levadas em conta na classificação do que chama grupo entoacional, embora julgue que a delimitação de um grupo entoacional deva ser feita *a priori* a partir de considerações fonéticas. A despeito dessa falta de consenso na definição de unidades entoacional, estudos têm mostrado que este é um conceito compreendido – ainda que intuitivamente – por especialistas e mesmo por

[4] Praat: doing phonetics by computer. Version 6.0.14, http://www.praat.org/

[5] Software by Edouard Geoffrois and Karim Boudahmane: http://transag.sourceforge.net/.

pessoas não treinadas, algo que é evidenciado por um alto grau de concordância entre anotadores (Buhmann *et al*. 2002; Mo *et al.* 2008; Wagner 2005).

Para assegurar a preservação dos arquivos digitais, estratégias de redundância e de backup foram cuidadosamente estudadas e implantadas. Backups incrementais programados foram realizados desde o início das atividades do projeto. Além disso, backups completos foram realizados com certa periodicidade em mídias físicas, guardadas em diferentes localidades, e em servidores diversos. Também foram estudados meios para o arquivamento dos dados digitalizados, de maneira a garantir a sua preservação para a posteridade, e para a disseminação dos mesmos, objetivando um alcance eficiente de todo o material do projeto às comunidades interessadas.

Os resultados dessas considerações metodológicas e as escolhas feitas a partir delas estão descritos no protocolo apresentado a seguir.

4. O Protocolo do Projeto NURC Digital

4.1 Digitalização

Nesta seção serão descritos os procedimentos utilizados na digitalização de arquivos de áudio e de arquivos de texto do Projeto NURC/Recife.

4.1.1 *Digitalização dos arquivos de áudio*

Para a digitalização dos arquivos de áudio do Projeto NURC/Recife, foram utilizados os seguintes equipamentos e acessórios:

- Gravador de rolo AKAI 4000 DS Mk-II, que possui uma cabeça de reprodução em linha de dois canais e quatro pistas, resposta de frequência de 30 a 22kHz, faixa dinâmica de +50dB e saída RCA com sensibilidade de 4 ± 1dB.
- Interface de áudio USB Sound Devices USBPre 2, que possui conversor A/D de 24 bits de resolução, faixa dinâmica de 114dB, resposta de frequência de 10 a 40kHz e entrada RCA com sensibilidade de +29 dBu min, +10 dBu max e impedância de 60kΩ.
- Cabo RCA Diamond JX-2055, com condutor de cobre enriquecido com prata, blindagem tripla, revestimento em PVC, conectores banhados a ouro 24K, insuladores de polietileno blindados por fita alumínio, impedância de 75Ω e revestimento externo em Nylon.

- Computador desktop iMac Intel Core i7 quad core de 27 polegadas, com 3,4GHz, Turbo Boost de até 3,9GHz, 32GB de memória embarcada SDRAM DDR3 - 1600MHz, armazenamento de 2TB.

A solução de software que melhor se adequou às necessidades do projeto e características dos arquivos de áudio a serem digitalizados correspondeu ao seguinte conjunto de programas:

- *Audacity*: programa gratuito, multiplataforma, de código aberto, que permite registro com taxa de amostragem de até 96kHz e 24 bits por amostra.
- *Audiofile Spectre*: analisador de áudio em tempo real, composto por diferentes medidores, tais como VU, LU, BBC, medidor de nível de entrada, medidor numérico com indicação de cortes de pico, etc. Trata-se de um programa pago, desenvolvido para o OS X.

Para o procedimento de digitalização, o gravador de rolo era conectado à interface de áudio USB através do cabo RCA. A interface de áudio, por sua vez, era conectada ao computador por cabo USB. O computador estava configurado para receber sinal de áudio da interface UBBPre.

Antes de iniciar a digitalização, as fitas de rolo eram rebobinadas duas vezes seguidas para evitar distorções mecânicas na reprodução, o que poderia ocorrer pelo fato de estarem guardadas e sem uso por um longo período de tempo, o que acarreta em desnivelamento da fita. Após o rebobinamento, os cabeçotes, guias e braços de pressão do gravador de rolo eram devidamente limpados com produto adequado, para garantir fidelidade máxima de reprodução da fita. Em seguida, realizavam-se testes em diversos níveis de volume com a gravação a ser digitalizada, com o objetivo de encontrar o nível de entrada ideal para a digitalização. Para isso, o ganho do amplificador da interface USB era regulado em um nível que permita uma digitalização em uma altura adequada, com um amplo *headroom*[6], sem causar sobrecarga de sinal e/ou cortes de pico. Por regra, um nível de -12 dBFS (decibéis relativos à escala completa) era observado para garantir uma relação sinal-ruído favorável e detalhes espectrais adequados. Quando ruídos constantes eram observados na gravação, aumentava-se a margem de *headroom*, diminuindo o nível de entrada de áudio para -18 dBFS.

[6] Margem de segurança: o espaço disponível entre o nível operacional de um sinal de áudio e o máximo permitido pelo equipamento. Por exemplo: uma gravação digital com nível médio de -15 dB que clipa em +20 dB tem 35 dB de *headroom*. Ou seja, o sinal pode aumentar até 30 dB sem que haja distorções.

A medição do nível de entrada era monitorada com o aplicativo *Spectre*, através de medidores PPM (*Peak Programme Meter*), que têm melhor resposta transitória que medidores VU (*Volume Unit*), muito comum em gravadores analógicos, como o equipamento AKAI 4000 DS Mk-II, que foi usado para reproduzir as fitas de rolo no processo de digitalização. A vantagem da utilização de um medidor de nível de entrada online neste caso se justifica não apenas pela possibilidade de uso de diferentes medidores, mas, sobretudo, pela possibilidade de registro dos níveis de entrada e de ocorrências indesejadas, como sobrecarga de sinal e cortes de pico, por exemplo. Como o processo de digitalização é longo, a dependência de inspeção manual contínua pode ser um complicador. Portanto, um registro histórico do nível de entrada é importante para garantir a qualidade dos dados digitalizados. Ao final do processo de digitalização, os registros históricos dos níveis de entrada eram considerados. Caso se observasse problemas relativos ao nível de entrada, a digitalização era refeita.

Como informado na seção anterior, todos os arquivos de áudio foram digitalizados com taxa de amostragem de 96.000 Hz e quantização de 24 bits. No *Audacity*, esta informação deve ser indicada na aba de qualidade dos arquivos, que se encontra no menu de preferências do aplicativo. Antes de salvar os arquivos, eles foram editados, de maneira a eliminar conteúdos não relevantes que antecediam ou procediam a gravação em si. No *Audacity*, os arquivos são salvos mediante a opção *exportar*. É importante observar aqui que o padrão do programa para arquivos WAV é de 16 bits. Portanto, é preciso selecionar em formatação, no processo de exportação, a opção "outros arquivos sem compressão", e então eleger, em *opções*, o formato WAV como *cabeçalho* e a opção Signed 24 bit PCM como *codificação*.

4.1.2 *Digitalização do material em papel*

O material em papel do Projeto NURC/Recife consistia em (i) transcrições das gravações, algumas datilografadas e muitas manuscritas; (ii) fichas com informações da gravação, do(s) informante(s) e do(s) documentador(es); (iii) miscelânea: documentos de administração, dados de análise, estudos quantitativos, etc.

A melhor solução encontrada para a digitalização do material em papel do Projeto NURC/Recife foi a seguinte:

- Scanner Epson GT-S55 Workforce Pro, com velocidade de até 25ppm/50ipm, possibilidade de digitalização frente e verso em uma só

passada, detecção automática de alimentação dupla, alimentador automático de grande capacidade e resolução ótica de 600 dpi.
- *ABBYY FineReader Professional*, programa usado para digitalizar e converter documentos para PDFs pesquisáveis, por meio de Reconhecimento Óptico de Caracteres (OCR), com uma precisão de até 99,8%, retendo a formatação e a estrutura multi-páginas dos documentos originais.

Todos os arquivos de papel foram escaneados em tons de cinza (*grayscale*), com resolução ótica de 300 dpi, seguindo recomendações de boas práticas para reconhecimento de caracteres (Tanner 2004). Muitas das folhas de papel estavam corroídas pela ação do tempo; outras, por serem muitos frágeis, não permitiam o processamento em conjunto com o alimentador automático do scanner, exigindo a digitalização folha por folha. Neste processo, muitas folhas rasgaram-se, e tiveram de ser recuperadas digitalmente. A depender da condição final do arquivo digitalizado, procedíamos à sua edição, de maneira a garantir um material mais próximo possível do original, sem marcas e falhas provenientes do processo de digitalização, com conteúdo pesquisável nos casos em que era possível realizar o reconhecimento de caracteres. Os arquivos foram, depois de processados e editados, salvos em formato PDF, seguindo recomendações de boas práticas para preservação de arquivos digitais de texto (Bandi *et al.* 2015; Hodge & Anderson 2007).

4.2 Anotação

Diferentemente da anotação original do Projeto NURC, que consistia basicamente de uma transcrição linear das gravações, o Projeto NURC Digital propõe uma anotação alinhada multi-nível dos dados. A seguir, apresentaremos o modelo de anotação do Projeto NURC Digital, adotado a partir de recomendações feitas por órgãos especializados em codificação de dados linguísticos e de considerações teórico-metodológicas norteadas pela especialidade e atuação dos coordenadores do projeto.

4.2.1 *Anotação manual*

A primeira etapa da elaboração do protocolo de anotação alinhada dos dados do projeto consistiu na eleição dos aplicativos que seriam utilizados como suporte para o processo. Após alguns testes, a escolha recaiu sobre os seguintes aplicativos:

- *Praat*, programa aberto, multiplataforma, gratuito, bastante utilizado em várias aplicações e análises linguísticas. O programa produz um arquivo de anotação chamado TextGrid, texto formatado (com estrutura de cabeçalho-corpo) contendo informações sobre o texto segmentado: tempo de início, tempo de fim e anotação correspondente a este intervalo temporal.
- *ELAN* (Wittenburg et al 2006), programa aberto, multiplataforma, gratuito, usado principalmente para anotação linguística. O programa produz um arquivo de anotação amplamente documentado, baseado em XML chamado EAF (ou EUDICO Annotation Format), contendo informações de metadados, de organização hierárquica da anotação e de anotação propriamente.

As transcrições das gravações do Projeto NURC, que servem como base para as demais anotações, foram segmentadas em unidades entoacionais, uma vez que são em geral consideradas unidades mais adequadas para a segmentação do discurso oral (Chafe 1994; Halliday 2004; Cruttenden 1986). A segmentação foi feita no aplicativo *Praat*, por anotadores devidamente treinados[7]. O *Praat* foi escolhido para o processo de segmentação por permitir uma melhor visualização do sinal acústico e, por conta disso, uma segmentação mais precisa. Portanto, a primeira linha de anotação, chamada *–ue* (unidade entoacional), é a base para as demais anotações. Optou-se, nesta linha de anotação, por uma transcrição ortográfica dos dados, por esta promover maior legibilidade para humanos e máquinas. Nesta transcrição, nenhum símbolo especial, formatação ou sinal de pontuação foi utilizado. Uma segunda linha de anotação, chamada *com* (comentário), traz informações gerais da gravação (como, por exemplo, presença de ruídos) e indica fenômenos para-linguísticos (riso, suspiro, elocução pausada, etc.) associados aos locutores na gravação. Em cada anotação associada a uma gravação haverá tantas linhas *–ue* quantos forem os participantes da gravação. Assim, por exemplo, em um inquérito do tipo DID, há as linhas *Doc-ue*, para a transcrição da fala do documentador e *Inf-ue*, para a transcrição da fala do informante. Opcionalmente, para os inquéritos do corpus compartilhado, foram incluídas, para cada linha *–ue*, uma linha correspondente com transcrições seguindo as normas originais do Projeto NURC, a título de ilustração. Esta linha foi chamada *–nurc*. A Figura 1 abaixo ilustra a anotação feita no *Praat*.

Todas as anotações foram revisadas três vezes, por diferentes anotadores, objetivando uma maior consistência da anotação.

[7] Testes de concordância foram realizados entre os anotadores antes do início da tarefa de segmentação. Todos indicaram um alto índice de concordância (Fleiss' kappa > 0.7).

4.2.2 *Anotação automática*

Para o processo de anotação automática, foi escolhido o *parser* PALAVRAS (Bick 2000). Trata-se de um *parser* que tem sido muito utilizado em vários corpora de fala do português brasileiro, como, por exemplo, o C-ORAL-BRASIL (Raso & Mello 2012), com alto índice de sucesso.

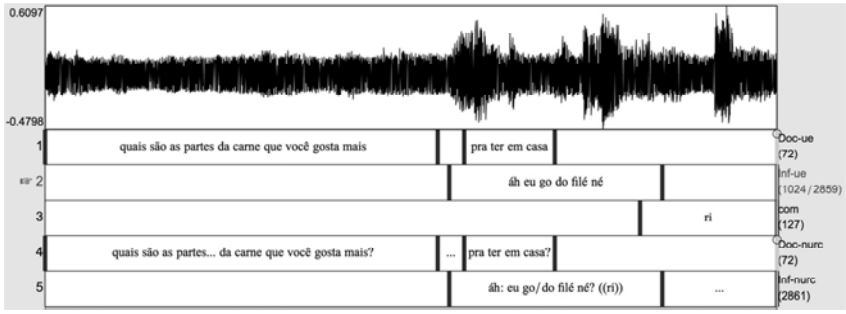

Figura 1. Exemplo de excerto do Inquérito NURC_RE_DID_150 anotado no *Praat*.

As anotações feitas no *Praat* foram individualmente processadas e editadas para gerar arquivos adequados para o processamento do *parser*. O *parser* PALAVRAS analisa morfossintaticamente cada palavra em um texto, colocando, para cada uma delas, marcações sobre sua classe morfológica e seu papel sintático. No entanto, para indicar o papel sintático das palavras no enunciado, o *parser* precisa de marcações específicas, que podem ser desde uma anotação prosódica até simplesmente sinais de pontuação. Nesse sentido, todas as linhas de anotação *–ue* foram salvas individualmente como arquivos de texto delimitado por tabulador no aplicativo *ELAN*. Neste arquivo, além da anotação propriamente, foram incluídas informações de tempo inicial e final de cada unidade entoacional. Os arquivos de texto resultantes ganharam marcas de pontuação, que foram revisadas por anotadores diferentes. Feito isso, os arquivos foram *tokenizados* no *ELAN* e alimentados no *parser* PALAVRAS.

O resultado da análise do *parser* foi importado na anotação manual, através do *ELAN*. Feito isso, os arquivos foram salvos no formato de anotação do *ELAN* (eaf) e no formato de anotação do *Praat* (TextGrid). A anotação completa dos dados do Projeto NURC consiste das seguintes linhas: (*-ue*) transcrição ortográfica dos enunciados segmentados em unidades entoacionais; (*-nurc*) transcrição segmentada em unidades entoacionais, seguindo as normas originais do Projeto NURC; (*com*) comentários; (*-plvs*) transcrição tokenizada em

palavras; (*lema*) lema ou lexia, a forma canônica da palavra; (*gram*) categoria gramatical, ou *part-of-speech* (POS); (*morf*) anotação morfológica; (*sint*) anotação sintática; e (*etk-2*) anotações secundárias (semântica, valência, subcategorias, etc). É importante notar que todas as linhas se repetem para cada linha –*ue*, exceto a linha *com*, que serve para o inquérito como um todo.

Figura 2. Exemplo de exceto do Inquérito NURC_RE_EF_259, com anotação completa no *ELAN*.

4.3 Arquivamento

Todos os dados do Projeto NURC estão organizados em uma estrutura hierárquica simples: o nó principal chama-se NURC_Digital. Nele encontra-se o sub-nó NURC_RE. Este sub-nó, por sua vez, divide-se em quatro sub-nós: NURC_RE_D2, NURC_RE_DID, NURC_RE_EF e NURC_RE_MISC. Dentro de cada um dos três primeiros nós há sub-nós correspondente a cada inquérito do NURC/RE. Os nós correspondentes aos inquéritos contém todos os arquivos referentes àquele inquérito: arquivo de áudio, ficha original com metadados (em pdf), arquivos de transcrição (em pdf) e arquivos de anotação (em TextGrid e eaf). O nó NURC_RE_MISC, por sua vez, contém arquivos de texto variados que faziam parte do acervo do Projeto NURC/RE (em pdf).

Cada nó e cada arquivo é acompanhado por metadados. Para a criação e edição de metadados, foi utilizado o aplicativo gratuito Arbil (Withers 2012). O aplicativo, feito em linguagem java, permite a criação de metadados tanto no formato *ISLE MetaData Initiative* (IMDI), quanto no formato *Component MetaData Infrastructure* (CMDI), ambos baseados na linguagem de marcação XML. O formato IMDI tem sido utilizado com maior frequência em projetos de documentação linguística e é o formato de escolha do arquivo selecionado para

o depósito do material do Projeto NURC Digital, o TLA (*The Language Archive*)[8].

É importante notar que o aplicativo permite selecionar que informações poderão ser anonimizadas. Trata-se de uma funcionalidade importante, uma vez que alguns dados de informantes (nomes, endereços, por exemplo) são sensíveis e, de acordo com as normas originais do Projeto NURC, devem ser protegidos.

Ao longo de todo o processo de digitalização e anotação de dados, diferentes estratégias de *backup* foram utilizadas, com o objetivo de resguardar o projeto de eventuais perdas dos arquivos originais, seja por ações despropositadas de usuários ou por mau funcionamento dos sistemas. A seguinte solução de *hardware* e *software* para *backup* foi utilizada:

- Disco rígido externo sem fio Airport Time Capsule 3TB, compatível com os computadores da Macintosh, que serve para fazer backups incrementais de maneira remota e automática de várias máquinas simultaneamente.
- Time Machine, programa de backup nativo do Mac OS, totalmente compatível com o Airport Time Capsule. Realiza, sob programação, backups incrementais periódicos.

Além de *backups* incrementais, foram realizados periodicamente *backups* completos em mídias óticas (ainda consideradas as mais duráveis), em um HD externo e no serviço de nuvens *Google Drive*. As mídias físicas eram mantidas sempre em lugares diferentes do local onde os dados primários estão custodiados, seguindo as recomendações de boas práticas de realização de cópias de segurança.

No que concerne o arquivamento dos dados do Projeto NURC Digital propriamente, este será feito no TLA, como mencionado acima, mediante a utilização do aplicativo LAMUS (Broeder *et al.* 2006). LAMUS (*Language Archive Management and Upload System*), assim como o Arbil, foi desenvolvido em formato java e funciona diretamente na internet. O programa permite que os usuários organizem e atualizem os conteúdos de um banco de dados e funciona a partir de metadados no formato IMDI. Em outras palavras: só permite o depósito de material associado a metadados, o que é uma exigência na maioria de arquivos digitais.

Por fim, o direito de acesso a determinados arquivos do corpus no TLA será administrado por meio do aplicativo baseado na internet *Access Management*

[8] https://tla.mpi.nl/.

System (AMS), desenvolvido pela equipe do TLA. Com este aplicativo, é possível restringir o acesso às fichas de metadados originais, não editadas, do Projeto NURC, que contém informações sensíveis.

4.4 Disseminação

Um dos objetivos centrais do Projeto NURC Digital foi disponibilizar os dados do Projeto NURC/Recife ao público em geral, através de um site dedicado. Esta ação beneficia diretamente a comunidade científica, que passa a ter disponíveis, para consulta otimizada, dados – anteriormente de difícil acesso – em formato digital de alta qualidade, devidamente catalogados, etiquetados e transcritos.

Nesse sentido, os dados foram tratados de maneira a adequá-los para diversos usos online, como busca avançada, opções de download, acesso a informações não-sensíveis, etc. Esse tratamento ocorreu em algumas etapas, a saber.

Todos os arquivos de áudio foram inspecionados e avaliados no que concerne à sua qualidade. Essa avaliação levou em consideração os seguintes critérios acústicos e perceptuais: a amplificação do sinal acústico, a uniformidade do sinal acústico, o volume, a inteligibilidade e a aceitabilidade[9]. Os dois primeiros critérios são mais objetivos, e serviram para identificar, no sinal acústico, a presença de cortes picos e de inconsistências nas variações de amplitude, que poderiam estar relacionados a artefatos do procedimento de digitalização. Os quatro últimos critérios são subjetivos e foram avaliados por julgadores treinados. A avaliação neste caso foi feita ao estilo do teste MOS (*Mean Opinion Score*), em uma escala de 1 a 5, em que 1 indica conceito mais negativo possível e 5 indica conceito mais positivo possível. Uma média global desses quatro últimos critérios serviu como valor final do MOS. O resultado da avaliação de todos os critérios serviu de base para algumas decisões referentes aos arquivos já digitalizados. A depender da seriedade dos problemas observados e do valor do MOS (< 2.5), a recomendação era de redigitalização. A recomendação para arquivos associados a um MOS entre 2.5 e 3.9 foi de melhoramento. Arquivos com um MOS superior a 4 eram considerados adequados. Os arquivos redigitalizados foram reavaliados. O valor de MOS está expresso nos metadados dos arquivos de áudio.

O processo de melhoramento dos arquivos de áudio digitalizados foi feito mediante a aplicação de técnicas de filtragens digitais que eliminam

[9] De acordo com Schmidt-Nielsen (1994) *aceitabilidade* no contexto de tecnologia da fala é uma medida subjetiva baseada no julgamento que ouvintes fazem acerca da qualidade do som de fala a que são expostos.

significativamente ruídos associados ao arquivo de som. Especificamente, o objetivo foi, principalmente, eliminar, ruídos *pitch* fixo – *hum* e *whistles* – que são comuns em gravações analógicas feitas em fitas magnéticas. Para o procedimento, foi utilizado o software Adobe Audition. Neste aplicativo, foi possível também eliminar cortes de pico associados à gravação analógica original e, a depender do caso, normalizar o volume, deixando-o mais adequado para usos que não requerem análises acústicas. Os arquivos melhorados estão disponíveis no site do Projeto NURC Digital, mas não foram arquivados. Apenas os arquivos originais, não-editados, foram arquivados.

Alguns arquivos de texto também foram editados, para mascarar informações sensíveis, como o nome dos informantes e os seus respectivos endereços. Este processo foi feito no aplicativo ABBYY FineReader Professional, que permite a edição de arquivos PDF.

Uma das ferramentas oferecidas no site do Projeto NURC Digital é o TEITOK, um sistema baseado na web para visualização, criação e edição de corpora anotados. O sistema fornece uma interface gráfica que possibilita a visualização da anotação de várias maneiras diferentes, a partir do interesse do usuário. Em particular, no caso do site do Projeto NURC Digital, o usuário pode, a partir de buscas específicas, ter acesso a exemplares de unidades entoacionais anotados contendo os ítens da busca, acompanhados de seus respectivos componentes sonoros, de forma rápida e eficiente. Para isso, o sistema faz uso de anotação feita em linguagem de marcação com padrões definidos pela *Text-Encoding Initiative* (TEI), indexados com codificação elaborada pelo *Corpus WorkBench* (CWB). Para tornar as anotações feitas pelo Projeto NURC Digital compatíveis com o TEITOK, foram convertidas para o formato XML, observando os padrões da TEI e as codificações do CWB. Este procedimento foi realizado de maneira semi-automática mediante um script Pearl desenvolvido pelo idealizador do TEITOK, Maarten Janssen. Metadados foram adicionados no cabeçalho manualmente após a conversão dos arquivos em XML.

As etapas descritas acima foram essenciais para a construção e alimentação o site do Projeto NURC Digital. O site, portanto, fornece os seguintes conteúdos: (i) informações gerais sobre o Projeto NURC e o Projeto NURC Digital; (ii) versões digitais dos volumes "Materiais Para o Seu Estudo", contendo transcrições do corpus compartilhado do Projeto NURC/Recife; (iii) todo o conteúdo do Projeto NURC/Recife, disponibilizado em estrutura hierárquica; (iv) corpus compartilhado totalmente anotado em dois formatos populares: TextGrid e eaf; (v) ferramenta de busca avançada por metadados para

localização de inquéritos que satisfaçam critérios específicos; (vi) sistema TEITOK, para visualização do corpus anotado, mediante busca avançada.

5. Considerações finais

O objetivo do presente artigo foi descrever os procedimentos que foram adotados pelo Projeto NURC Digital para a digitalização, a anotação, o arquivamento e a disseminação do material do Projeto NURC/Recife, que resultaram em um protocolo de boas práticas a serem adotadas para a digitalização de dados linguísticos gravados em formato analógico, em geral, e, em particular, para a digitalização dos dados do Projeto NURC; e para a anotação, o arquivamento e a disseminação de dados linguísticos em formato digital. Este protocolo, com detalhes de cada uma das etapas que o caracteriza, está disponível para consulta, sob solicitação.

O processo de digitalização dos dados do NURC/Recife evidenciou a fragilidade com que se encontra esse material. Dada a riqueza dos dados do Projeto NURC e a incontestável importância que esse material teve e tem para a pesquisa em linguística desenvolvida no Brasil, torna-se imperativo a adoção de práticas urgentes para garantir a sua preservação nas demais capitais em que o projeto foi sediado: Salvador, Rio de Janeiro, São Paulo e Porto Alegre. Esperamos que este artigo sirva como estímulo para isso.

Agradecimentos

Este trabalho não seria possível sem a colaboração dos seguintes pesquisadores, em diferentes etapas do projeto: Dóris de Arruda C. da Cunha (UFPE), Marcos Galindo Lima (UFPE) Paul Trilsbeek (MPI, Nijmegen), Eckhard Bick (SDU, Odense), Maarten Janssen (UL, Lisboa). De fundamental importância foi a colaboração de pesquisadores de iniciação científica ao longo do desenvolvimento das atividades do projeto: Ebson Silva, Tiberio Correia, Remildo Silva, Ingrid Rodrigues, Maxswell Santos, Clara Silva, Nayara Leite, Julyana Silva, Dieggo Arnoldo, Vinícius Pereira, Juliete Melo, Maristela Santos e Remildo Silva. Por fim, o autor deseja agradecer o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que financiou o projeto NURC Digital, através do Edital Universal 14/2012, Processo no 472918/2012-5.

Referências

Abaurre M.B.M. & Rodrigues, Â.C.S. (eds.) 2002. *Gramática do português falado*. Campinas: Editora da Unicamp.

Bandi, S., Angadi, M. & Shivarama, J. 2015. Best Practices in Digitization: Planning and Workflow Processes. In M. Angadi, G.Z. Shinde, P.S. Kattimani & S. Jange (eds), *Emerging Technologies and Future of Libraries Issues and Challenges*. New Delhi: Daya Publishing House, 332-339.

Bick, E. 2000. *The parsing system "PALAVRAS": automatic grammatical analysis of Portuguese in a constraint grammar framework*. Aarhus: Aarhus University Press.

Bradley, K. (ed.) 2009. *Guidelines on the Production and Preservation of Digital Audio Objects*. Australia: International Association of Sound and Audiovisual Archives.

Broeder, D., Claus, A., Offenga, F., Skiba, R., Trilsbeek, P. & Wittenburg, P. 2006. LAMUS: The Language Archive Management and Upload System. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*. http://www.lrec-conf.org/proceedings/lrec2006/ (accessed November 24, 2016).

Burg, J., Romney, J. & Schwartz, E. 2016. *Digital Sound & Music. Concepts, Applications, and Science*. Portland: Franklin, Beedle & Associates Inc.

Buhmann, J., Caspers, J., van Heuven, V.J., Hoekstra, H., Martens, J.P. & Swerts, M. 2002. Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. In *Proceedings of LREC 2002, Third International Conference on Language Resources and Evaluation*. http://www.lrec-conf.org/proceedings/lrec2002/ (accessed November 24, 2016).

Callou, D.M.I. (ed.). 1992. *A Linguagem Falada Culta na Cidade do Rio de Janeiro. Materiais para seu estudo. Vol. I - Elocuções Formais*. Rio de Janeiro: UFRJ/FJB.

Callou, D.M.I. & Lopes, C.R. (eds.). 1993. *A Linguagem Falada Culta na Cidade do Rio de Janeiro. Materiais para seu estudo. Vol. II - Diálogo entre Informante e Documentador*. Rio de Janeiro: UFRJ/CAPES.

Callou, D.M.I. & Lopes, C.R. (eds). 1994. *A Linguagem Falada Culta na Cidade do Rio de Janeiro.Materiais para seu estudo. Vol. III - Diálogos entre dois informantes.* Rio de Janeiro: UFRJ/CAPES.

Castilho, A. (ed.). 1990. *Gramática do português falado*. Campinas: Editora da Unicamp; São Paulo: Fapesp.

Castilho, A. (ed.). 1993. *Gramática do português falado*. Campinas: Editora da Unicamp; São Paulo: Fapesp.

Castilho, A. 2007. Fundamentos teóricos da Gramática do português culto falado no Brasil. *Alfa* 51(1): 99-135.

Castilho, A. & Basílio, M. (eds.). 1996. *Gramática do português falado*. Campinas: Editora da Unicamp; São Paulo: Fapesp.

Castilho, A., Oliveira E Silva, G.M.D. & Lucchesi, D. 1995. Informatização de acervos da Língua Portuguesa. *Boletim da Associação Brasileira de Lingüística* 17: 143-154.

Castilho, A. & Preti, D. (eds). 1986. *A Linguagem Falada Culta na Cidade de São Paulo. Materiais para seu estudo. Vol. I - Elocuções Formais*. São Paulo: TAQ/Fapesp.

Castilho, A. & Preti, D. (eds). 1987. *A Linguagem Falada Culta na Cidade de São Paulo. Materais para seu estudo. Vol. II - Diálogos entre dois informantes*. São Paulo: TAQ/Fapesp.

Chafe, W.L. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: Chicago University Press.

Cruttenden, A. 1986. *Intonation*. Cambridge: Cambridge University Press.

Crystal, D. 1969. *Prosodic Systems and Intonation in English*. London: Cambridge University Press.

da Cunha, D.A.C. 2003. A produção de sentido na fala e na escrita. *Revista do GELNE* 3: 27-32.

Edwards, J.A. & Lampert, M.D. (eds). 1992. *Talking data: Transcription and coding in discourse research.* Hillsdale: Lawrence Erlbaum Associates.

Halliday, M.A.K. 2004. *An Introduction to Functional Grammar*. London: Edward Arnold.

Hilgert, J.G. (ed.). 1997. *A Linguagem Falada Culta na Cidade de Porto Alegre. Vol. I - Diálogos entre informante e documentador*. Passo Fundo: Ediupf; Porto Alegre: Ed. Universidade/Ufrgs.

Himmelmann, N.P. 2006. Language documentation: what is it and what is it good for? In J. Gippert, N.P. Himmelmann & U. Mosel (eds). *Essentials of Language Documentation*, Berlin: Mouton de Gruyter, 1-30.

Hodge, G. & Anderson, N. 2007. Formats for Digital Preservation: A review of alternatives and issues. *Information Services & Use* 27(1-2): 45-63.

Ilari, R. (ed.). 1992. *Gramática do português falado: níveis de análise linguística*. Campinas: Editora da Unicamp.

Kato, M. (ed.). 1996. *Gramática do português falado: convergências*. Campinas: Editora da Unicamp; São Paulo: Fapesp.

Koch, I.G.V. (ed.). 1996. *Gramática do português falado*. Campinas: Editora da Unicamp; São Paulo: Fapesp.

Lieberman, P. 1967. *Intonation, Perception and Language*. Cambridge: MIT Press.

MacWhinney, B. 2000. *The CHILDES project: tools for analyzing talk*. Mahwah: Lawrence Erlbaum.

Marcuschi, L.A. 1986. *Análise da conversação*. São Paulo: Ática, 1986.

Martin, P. 2004. WinPitch Corpus: a text to speech alignment tool for multimodal corpora. In M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa & R. Silva, *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*. Paris: ELRA.

Mo, Y., Cole, J., & Lee, E.K. 2008. Naive listeners prominence and boundary perception. In P.A. Barbosa, S. Madureira & C. Reis (eds), *Proceedings of the Speech Prosody 2008 Conference*. Campinas: Universidade de Campinas.

Motta, J. & Rollemberg, V. (eds). 1994. *A Linguagem Falada Culta na Cidade de Salvador. Materiais para seu estudo. Vol. I - Diálogos entre Informante e Documentador*. Salvador: Instituto de Letras da UFBa.

de Moura Neves, M.H. (ed.) 1999. *Gramática do português falado*. São Paulo: Humanitas; Campinas: Editora da Unicamp.

Oliveira, M. 1999. The Function of Self-Aggrandizement in Storytelling. *Narrative Inquiry*, 9(1): 25-47.

Paiva, M.C. 2003. Transcrição de dados linguísticos. In M.C. Mollica & M.L. Braga (eds), *Introdução à Sociolinguística: o tratamento da variação*. São Paulo: Contexto, 135-146.

Percy, C.E., Meyer C.F., Lancashire, I. (eds.). 1996. *Synchronic Corpus Linguistics. Papers from the sixteenth International Conference on English Language and Research on Computerized Corpora (ICAME 16)*. Amsterdam/Atlanta: Rodipi.

Plichta, B. & Kornbluh, M. 2002. *Digitizing Speech Recordings for Archival Purposes*. http://www.historicalvoices.org/papers/audio_digitization.pdf (accessed November 24, 2016).

Preti, D. & Urbano, H. (eds). 1990. *A Linguagem Falada Culta na Cidade de São Paulo. Materiais para seu estudo. vol. III - Diálogos entre o Informante e o Documentador*. São Paulo: TAQ/Fapesp.

Raso, T. & Mello, H. (eds) 2012. *C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Universidade Federale de Minas Gerais.

Ramilo, M.C. & Freitas, T. 2010. Transcrição Ortográfica de Textos Orais: Problemas e Perspectivas. In Oliveira, M. (ed.). *Estudos de Corpora. Da teoria à prática*. Lisboa: Colibri, 67- 83.

Sá, M.P.M. 2004. Estrutura e natureza da narrativa na conversação. *Boletim Informativo* 32. Maceió: UFAL. Apresentado no *XIX Encontro Nacional da Anpoll*, na Universidade Federal de Alagoas.

Sá, M.P.M., da Cunha, D.A.C., Lima, A.M. & Oliveira, M. (eds). 1996. *A Linguagem Falada Culta na Cidade do Recife. Vol. I - Diálogos entre informante e documentador*. Recife: Universidade Federal de Pernambuco.

Sá, M.P.M., da Cunha, D.A.C., Lima, A.M. & Oliveira, M. (eds). 2005. *A linguagem falada culta na cidade do Recife: elocuções formais*. Recife: Universidade Federal de Pernambuco.

Sacks, H., Schegloff, E. & Jefferson, G. 1978. A simplest systematics for the organization of turn taking for conversation. In J. Schenkein (ed.), *Studies in the Organization of Conversational Interaction*. New York: Academic Press, 7-56.

Sardinha, T.B. 2000. Linguística de Corpus: Histórico e Problemática. *D.E.L.T.A.* 16(2): 323-367.

Selkirk, E.O. 1984. *Phonology and Syntax: The Relation Between Sound and Structure*. Cambridge: MIT Press.

Schmidt-Nielsen, A. 1994. Intelligibility and Acceptability testing for Speech Technology. In A.K. Syrdal, R.W. Bennett, S.L. Greenspan (eds), *Applied Speech Technology*. Boca Raton: CRC Press.

Tanner, S. 2004. *Deciding Whether Optical Character Recognition is Feasible*. Published by King's Digital Consultancy Services. http://www.odl.ox.ac.uk/papers/OCRFeasibility_final.pdf (accessed February 20, 2016).

Von Arb, J. & Gaustad, L. 2005. Guidelines on the Production and Preservation of Digital Audio Objects – optimizing quality access through digital preservation practice. In *World Library and Information Congress: 71th IFLA General Conference and Council*. Oslo, Norway.

Wagner, P. 2005. Great Expectations – Introspective vs. Perceptual Prominence Ratings and their Acoustic Correlates. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*. Lisbon: ISCA, 2381–2384.

Withers, P. 2012. Metadata Management with Arbil. In V. Arranz, D. Broeder, B. Gaiffe, M. Gavrilidou & M. Monachini (eds), *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation in the Documentation of LR at LREC 2012*. Istanbul: ELRA, 72-75.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. & Sloetjes, H. 2006. ELAN: a Professional Framework for Multimodality Research. In *Proceedings of LREC 2006,*

Fifth International Conference on Language Resources and Evaluation. http://www.lrec-conf.org/proceedings/lrec2006/ (accessed November 24, 2016).

Van Bogart, J. 1995. *Magnetic Tape Storage and Handling: A Guide for Libraries and Archives*. Washington, D.C.: Commission on Preservation and Access. http://www.clir.org/pubs/reports/pub54/Download/pub54.pdf (accessed February 20, 2016).