PARA A ELABORAÇÃO DO BANCO DE DADOS NEUROLINGÜÍSTICOS - BDN
(FOR THE BUILDING OF THE NEUROLINGUISTIC DATA ARCHIVE – BDN)

Cilene CAMPETELA (UNICAMP)

ABSTRACT: This paper is to present the process of building of the BDN *– a neurolingüistic data archive; and categories of search code. In this context, it is using descriptive categories instead of analitics categories, by the fact of the researchers, and not the group of transcription, have to analyse the* BDN corpora.

KEYWORDS: neurolingüistics, data, linguistic categories.

0. Introdução

Esta comunicação visa apresentar o processo de formação do Banco de Dados de Neurolingüística (BDN), constituídos a partir da degravação de dados do Centro de Convivência de Afásicos - CCA/IEL/FCM - (registrados desde 1989). Apresentaremos uma visão de conjunto do trabalho de transcrição e composição do BDN, desde 1992, quando do primeiro período de vigência deste Projeto Integrado. Explicitaremos a situação atual desse trabalho destacando as questões e avanços referentes ao período de 1998 a 1999, quando começamos a formular o Código de Busca do BDN e ocorreram algumas avaliações por parte da Equipe de Transcrição.

A formação do Banco de Dados de Neurolingüística (BDN) é um dos objetivos do Projeto Integrado de Pesquisa "Contribuições da pesquisa neurolingüística para a avaliação do discurso verbal e não-verbal" (CNPq: 521773/95-4). Nesse sentido é crucial a participação de outros pesquisadores (como tivemos uma amostra em dezembro de 1998 com a jornada sobre Banco de Dados em Lingüística, justamente para discutir e conhecer a experiência do banco de dados do VARSUL e do CHILDES apresentada pelas profª^{as} Drª^{as} Iara Bemquerer Costa e Ester Scarpa, respectivamente).

Em maio de 1999, dispusemos um fac-símile de quatro sessões de 1990 no formato atual do BDN aos pesquisadores do projeto, tendo dois objetivos: ter sua colaboração na formulação das categorias do sistema de busca, apresentadas nesta comunicação e avaliar a eficácia do sistema de busca (Word) que estamos utilizando.

As sessões do CCA (que se reúne duas vezes por semana durante duas horas) são gravadas e depois transcritas para que componham o BDN. O processo de formação deste Banco tem sido evolutivo, ou seja, sua implantação tem passado por etapas significativas de adequação e modernização: a) adequação aos interesses da pesquisa neurolingüística, tornando o BDN ainda mais acessível ao desenvolvimento de pesquisas sob vários enfoques: fonológico, sintático, discursivo *etc*.; b) modernização, para que os dados possam ser rapidamente selecionados através de um programa de computador, o que possibilita uma coleta exaustiva de informações pertinentes ao

desenvolvimento de pesquisas e, além disso, dispensa o trabalho mecânico de seleção de dados, economizando-se este tempo de seleção para aplicá-lo na análise propriamente dita do fenômeno que estiver em questão.

1. Levantamento diacrônico do trabalho de elaboração do BDN:

Entre 1990 e 1994, o BDN consistia num acervo composto por fitas cassete gravadas durante as sessões do CCA-I, que eram transcritas por alunas do curso de lingüística e por fonoaudiólogas, todas dedicadas à pesquisa, ora no nível de Iniciação Científica, ora no de Aperfeiçoamento, ora no de Pós-Graduação, acompanhando e observando os processos evidenciados pelos quadros patológicos dos pacientes que compunham o grupo.

Esta primeira fase foi muito importante para a formação do BDN. No entanto, as pesquisadoras sentiam necessidade do registro visual das sessões, além do registro auditivo, tendo em vista que este tipo de registro não contemplava fatos e pesquisas de natureza não verbal.

Para que fosse formado o *corpus* de análise do discurso não-verbal, tendo em vista a importância da atividade gestual para a pesquisa neurolingüística e considerando que a praxia é um fato extremamente relevante no contexto patológico das afasias, os pesquisadores deviam participar de todas as sessões e elaborar registros escritos a partir da observação dos recursos não-verbais utilizados pelos pacientes. Por mais que o registro por escrito fosse eficiente, a descrição dos gestos é diferente de sua forma de ocorrência. Por outro lado, não havia, até então, como subsidiar as análises da atividade gestual e sua relação com a linguagem, principalmente em situações de divulgação (congressos, palestras, aulas, artigos *etc*.) por meio da exibição dos dados coletados. O registro por escrito revelou-se também de difícil consulta no momento de transcrever, o que tornava muito mais moroso o processo do que se previu.

A partir de outubro de 1995 (06 sessões de duas horas cada), com o interesse de adequar o BDN às necessidades da pesquisa, as sessões do CCA, grupo I, passaram a ser filmadas. Este novo processo de formação do BDN foi um grande passo para o aprimoramento das condições de registro dos dados da área, abrindo novas frentes de pesquisa sobre dados de oralidade, e com múltiplos interlocutores (afásicos e não afásicos). Os textos produzidos por Luciana Sales Pires (1998) relatam bem a metodologia de transcrição que conseguimos desenvolver de modo a relacionar os dois modos de registro - áudio e vídeo - em benefício dos dados que compõem o BDN.

Desta possibilidade que se abriu, pesquisadores estudiosos de Artes Cênicas manifestaram interesse de contribuir para a "análise do gesto e de sua dimensão simbólica quando do funcionamento da linguagem". Eles passaram a compor o quadro de pesquisadores do Projeto Integrado em Neurolingüística, em 1996, desenvolvendo um programa de Expressão Teatral junto aos dois grupos do CCA.

Em relação ao registro de dados do CCA, percebeu-se, também, que a gravação em vídeo não poderia substituir a gravação em fitas cassete, porque a qualidade auditiva do vídeo não é tão eficiente quanto a do áudio, o que poderia prejudicar enormemente a transcrição dos dados. Por este motivo, as sessões dos dois

grupos do CCA, em 1996, passaram a ser gravadas das duas formas: em fitas de vídeo e em fitas cassete e a equipe de transcrição investiu na metodologia de transcrição aúdio e vídeo.

Como já foi dito, a transcrição das fitas com dados coletados durante as sessões do CCA-I eram transcritas por uma equipe de alunas (IC) que se integravam em pesquisas neurolingüísticas.

No início, a transcrição era feita em programas de computadores, mas não seguiam uma regulamentação, o que resultou num acervo composto por transcrições feitas nos mais variados programas de computador, tipos e tamanhos distintos de papel (carta, A4, tamanho ofício etc.), além de vários tipos de programas (wordstar, courrier, times new roman, arial) e tamanhos distintos de letras (12, 14, padrão do wordstar). Toda essa variedade se dava por falta de recursos materiais para subsidiar a transcrição. Talvez este problema de diversidade dos tipos de transcrição não fosse tão relevante se não fosse a necessidade de se informatizar o BDN e facilitar a seleção dos dados.

Para que a metodologia de transcrição fosse unificada, houve a elaboração de um primeiro conjunto de regras de codificação, que passaram a ser seguidas por cada uma das transcritoras da equipe.

A elaboração destas regras foi um processo complexo, porque dependia da adequação da transcrição aos programas de computador que passaram, também, por várias modificações: do Windows 3.1 ao Windows 3.11 para Workgroups e, logo em seguida, para o Windows 95 que pode ser "lido" pelo Windows 97, também. Além disso, passou-se por várias fases do Word for Windows, estabilizado na última versão do Windows, que "lê" todas as versões anteriores.

A transcrição continuou, mas algumas transcritoras passaram a depender dos computadores da Universidade, já que não dispunham de um programa compatível com as novas regras. Iniciou-se, portanto, uma série de perdas de arquivos por contaminação por vírus e, também, várias vezes os computadores não estavam liberados ou estavam ocupados. A situação dos computadores não está normalizada até hoje.

Diante dos problemas e das novas regras de codificação, a equipe passou a ser dividida em quatro transcritoras e uma revisora das transcrições antigas, tendo em vista que o trabalho de transcrição deveria continuar, ao mesmo tempo que se revisasse, adequadamente e atualizasse o existente. A revisão ficou a cargo do Apoio Técnico, concedido pelo FAEP/UNICAMP, em 1996, e quem trabalhou nesse processo foi a doutoranda em Lingüística Cilene Campetela (IEL/UNICAMP).

Foram notados, neste processo, alguns problemas de adequação da transcrição, tendo em vista que os dados não forneciam subsídios para a realização de trabalhos fonéticos e prosódicos por não haver transcrição fonética destes dados. É de se observar que a transcrição destes fatos de linguagem patológica é extremamente complexa de ser registrada e requer uma formação técnica bastante especializada.

Ao mesmo tempo que o sistema de transcrição parecia dar conta de registrar as sessões do CCA-I e construir um bom acervo que proporcionasse aos pesquisadores a seleção de dados pertinentes às várias pesquisas, ele era ineficiente para o desenvolvimento de estudos sobre a interferência da afasia nos processos fonético-fonológicos da produção verbal. Este sistema não dispunha de registros capazes de

diferenciar, por exemplo, afasias afeitas ao nível fonológico do semântico, em termos das parafasias (relacionadas ao nível de linguagem e à repercussão deste no funcionamento discursivo da linguagem) que afetam a produção verbal e processos interpretativos constitutivos desses quadros, bem como os recursos verbais utilizados pelos sujeitos afásicos para suprir tais dificuldades.

A transcrição era, pois, ortográfica e, na presença de algum desvio por deformação do trato oral, por exemplo, em que não era possível processar através da transcrição ortográfica a produção de fala de determinados pacientes, as lacunas eram preenchidas com a expressão trecho ininteligível.

Passou-se, então, a realizar várias reuniões em que se discutia sobre a necessidade de transcrever foneticamente alguns dos dados do BDN, a fim de que eles pudessem ser eficientes também para a pesquisa dos sons produzidos e da sua interface com a semântica, já que a grande maioria das produções orais eram compreendidas, bem engajadas no discurso, apesar de não serem bem produzidas, do ponto de vista fonético.

Para a implantação da transcrição fonética dos dados do BDN houve o início de uma nova revisão, desta vez minuciosa, para que todas as transcrições, desde 1990, pudessem conter as mesmas informações e, com isso, o BDN estaria mais adequado às pesquisas lingüísticas.

Ainda assim, a transcrição fonética muitas vezes não dava conta do sentido do enunciado; por este motivo, e com base no artigo - A importância do tom para a compreensão da linguagem (Campetela, 1996) - em que a autora explicita a importância do estudo prosódico, introduzimos novas mudanças no tratamento dos contextos verbais em que a transcrição fonética é necessária, incluindo também a transcrição prosódica da entonação de cada sujeito afásico.

O trabalho que inicialmente se fundamentava pela transcrição ortográfica passou a ser um trabalho lingüisticamente mais técnico. Por este motivo, a equipe foi reorganizada, substituindo-se as quatro transcritoras anteriores por outras que realizassem não só uma transcrição ortográfica, como também se empenhassem na aplicação dos seus conhecimentos de fonética e fonologia aos dados do BDN do CCA-I.

O processo de (re)formulação do BDN não parou por aqui. Hoje, pode-se contar com uma Macro exclusiva, criada no Setor de Apoio à Informática do IEL/UNICAMP, que é capaz de realizar, rapidamente, através do Word, uma busca de dados, fazendo com que os dados pertinentes a quaisquer pesquisas em Neurolingüística possam ser listados de acordo com o código de busca selecionado pelo pesquisador.

2. A Formação do Código de Busca:

O Código de Busca faz parte da quarta etapa de formação do BDN, momento em que já estão satisfatoriamente consolidadas as regras de transcrição e o formato mais adequado para o armazenamento dos dados neurolingüísticos.

A atual versão do BDN propicia agilidade na seleção de dados pertinentes às pesquisas em Neurolingüística. Estes dados, independentemente do código de busca

estabelecido, estarão sempre disponíveis para que os pesquisadores os consultem também através de cópias impressas e das próprias fitas.

A Macro exclusiva foi idealizada por Cilene Campetela e elaborada por Sueli Rizzoli, do Setor de Apoio à Informática do IEL/UNICAMP.

No princípio, o código de busca vinha sendo implantado pela equipe de transcrição, baseando-se em informações que posicionassem cada enunciado do discurso de acordo com três níveis: 1) Atos de Fala; 2) Informações empíricas do discurso; 3) Informações prosódicas (entonacionais) do discurso. Já neste momento da formulação, o cruzamento entre os níveis podia ser "lido" pela Macro, eliminando-se desta forma qualquer problema que pudesse existir com relação à escolha do código de busca que possuía as seguintes entradas, baseadas em categorias de análise:

ATOS DE FALA	FATOS DE ORALIDADE	INFORMAÇÕES PROSÓDICAS
\ ? : enunciado interrogativo	\ hes: hesitação	\ TF: transcrição fonética
\ neg: enunciado negativo	\ né:	\ tom: intonação
\ top: topicalização	\ tá:	\ : : alongamento vocálico
\ imp: imperativo	\ ins: inserção	\ /: pausa breve
	\ rep: repetição	\ //: pausa longa
	\ pars: parafasia semântica	
	\ parf: parafasia fonológica	

Este código de busca não foi, entretanto, considerado adequado porque cabia à equipe de transcritoras do BDN realizar análises um tanto quanto especulativas dos dados. Visto que o objetivo da transcrição deve se limitar à descrição dos dados para que os pesquisadores desenvolvam suas análises, a Equipe do BDN, juntamente com a coordenadora e outros pesquisadores do Projeto Integrado, reavaliaram e reformularam as categorias da seguinte forma:

CÓDIGO	FINALIDADE
\tom	Entonação utilizada pelo falante no momento do enunciado ininteligível.
\TF	Transcrição fonética
\hes	Hesitação.
\ top	topicalização sintática
\ neg	enunciado negativo
\ ins	inserção
\ aí	Aí, daí, então
\ né	
\ tá	
\ rir	risos
\ int	introdução de Topoi ("eu acho que")
\ lei	leitura em voz alta

\ com	comparação (“mais que”, menos que”)
\ esc	escrita

Com este novo Código de Busca, os dados são descritos pela equipe de transcrição de forma que possam ser agrupados genericamente, permitindo que, a partir das categorias selecionadas, os pesquisadores obtenham somente dados relevantes para suas análises nas mais diversas áreas da Lingüística: fonética e fonologia, sintaxe, semântica, análise do discurso *etc*.

Como é possível observar, o BDN sofreu vários contratempos que resultaram em dificuldades até que os ajustes e reformulações fossem nele aplicados de forma melhor acabada. Ao contrário do que se possa imaginar, o trabalho passou a ser muito mais complexo a ponto de recobrir as pesquisas que se desenvolvem na área de Neurolingüística propiciando, na verdade, uma melhor adequação e modernização do BDN do CCA, a ser aplicado a todos os dados da área.

Além disso, atualmente, o BDN se encontra muito melhor elaborado, permitindo uma maior velocidade às pesquisas que se realizam através dele, além de possibilitar o acesso de pesquisadores de várias áreas afins.

RESUMO: Este artigo pretende mostrar o processo de formação do BDN – Banco de Dados Neurolingüísticos; e as categorias do código de busca. Neste sentido, utilizamos categorias descritivas em vez de categorias de análise porque apenas os pesquisadores, e não o grupo de transcrição, deve fazer a análise dos fenômenos encontrados neste corpus que compõe o BDN.

PALAVRAS-CHAVE: neurolingüística, dados, categorias lingüísticas.

REFERÊNCIAS BIBLIOGRÁFICAS:

CAMPETELA, C. (1994) A importância do tom para a compreensão da linguagem. I Congresso Internacional da ABRALIN. Salvador/BA.

_____ (1998) Descrição diacrônica do processo de formação do banco de dados neurolingüísticos – BDN. Jornada de Banco de Dados em Lingüística. UNICAMP.

_____ (1999) Para um Sistema de Codificação dos Dados Neurolingüísticos. Qualificação de Área apresentada como requisito para defesa de tese em nível de Doutorado.

PERONI, L. S. (1998) Transcrição de Áudio de Dados do Centro de Convivência – CCA. VI Congresso Interno de Iniciação Científica da UNICAMP e Jornada de Banco de Dados em Lingüística (1999). UNICAMP

PIRES, L. S. (1999) Degravação de Vídeo em Neurolingüística. PIBIC.

_____ (1999) Degravação em Vídeo para a Pesquisa Neurolingüística. Jornada de Banco de Dados em Lingüística. UNICAMP.