Artefactum
Revista de estudos em Linguagens e Tecnologia

O CORPUS *NALINGUA* E AS TECNOLOGIAS DE APOIO: A CONSTITUIÇÃO DE UM BANCO DE DADOS DE FALA DE CRIANÇAS NO BRASIL

Alessandra Del Ré
aledelre@fclar.unesp.br
http://lattes.cnpq.br/4188483346917822
Rosângela Nogarini Hilário
ronogarini@gmail.com
http://lattes.cnpq.br/0552571325354423
Rúbens Antonio Rodrigues
rubensan.rodrigues@gmail.com
http://lattes.cnpq.br/7493315125981109

RESUMO

Este artigo pretende demonstrar de que forma a tecnologia pode contribuir para os estudos que envolvem a linguagem, no caso, a linguagem da criança. Para tanto, trazemos questões mais genéricas relativas à constituição de um banco de dados de fala de crianças, e outras mais específicas pertencentes ao *corpus* do Grupo NALingua (CNPq). Inicialmente, discutem-se dificuldades metodológicas que interferem diretamente nas pesquisas da área de Aquisição da Linguagem, sobretudo no que se refere à coleta, transcrição e compartilhamento de dados; em seguida, apresentam-se soluções tecnológicas práticas para a difícil tarefa de constituição de um banco de dados que se pretende disponibilizar à comunidade acadêmica.

Palavras-chave: aquisição da linguagem; metodologia de pesquisa; novas tecnologias.

Introdução

O objetivo deste artigo é mostrar de que forma a tecnologia pode contribuir para estudos que envolvem a linguagem, no caso, a linguagem da criança.

As pesquisas que se dedicam ao trabalho com dados de fala da criança e inserem-se na área de Aquisição da Linguagem – uma subárea da (Psico)Linguística – envolvem estudos sobre a aquisição oral de língua materna (em crianças com e sem comprometimento de linguagem), aquisição de segunda língua/aprendizagem de LE e aquisição da escrita (DEL RÉ, 2006).

Uma questão cara à área, pelas dificuldades que ela impõe, é a metodologia, que a cada pesquisa necessita ser repensada em função da perspectiva teórica, dos objetivos etc (HILÁRIO, DEL RÉ, 2015). A coleta de dados e seu armazenamento também são

preocupações constantes. No início, antes mesmo da Psicolinguística se estabelecer enquanto tal em meados da década de 1950, eram feitos diários com dados de fala das crianças, anotações em geral feitas pelos próprios pais. Mas como discernir realmente o que a criança produziu, já que, nesse caso, o dado vem sempre carregado da interpretação de alguém que o relata? Com a tecnologia, surgiram gravações em áudio, em seguida em vídeo com o objetivo de tornar esses dados mais fidedignos. Essa tecnologia foi se aperfeiçoando, com maior qualidade de imagem, som, programas de computador, e com transcrições que permitem ao analista, com um clique no enunciado transcrito, ver e/ou escutar o vídeo e/ou som correspondentes. É a partir dessa nova realidade que o *corpus* do grupo de pesquisa NALingua (CNPq) [1], Núcleo de Estudos em Aquisição da Linguagem, está se constituindo.

Tal grupo é composto por uma equipe interdisciplinar, que reúne linguistas, psicólogos, fonoaudiólogos, educadores, do Brasil[2], em colaboração com pesquisadores da França[3], e que se propõem a analisar um mesmo conjunto de dados, com recortes específicos dos objetos de análise, visando apreender o processo de constituição da criança enquanto sujeito falante, desde as primeiras vocalizações, elementos prosódicos, gestos, até a entrada na escrita. Trata-se de dados de crianças monolíngues e bilíngues, com perfil socioeconômico de classe média e classe média alta, registrados preferencialmente em ambiente familiar, naturalístico, com a presença dos pais. O conjunto de dados é composto por:

---

[1] dgp.cnpq.br/dgp/espelhogrupo/4240032996711008

[2] Alessandra Del Ré (coord.) – UNESP/Linguística; Márcia Romero (vice-coordenadora) – UNIFESP/Educação; Marianne Carvalho Bezerra Cavalcante – UFPB/Linguística; Carmem Luci da Costa Silva – UFRGS/Linguística; Claudemir Belintane – USP/Educação; Eduardo Calil – UFAL/Educação; Gladis Massini-Cagliari – UNESP/Linguística; Irani Maldonade – UNICAMP/Fonoaudiologia; Marly Matos – USP/Letras Clássicas; Selma Leitão – UFPE/Psicologia Cognitiva; Zelita F. Guedes – UNIFESP/Fonoaudiologia; Christelle Dodane – Université de Montpellier 3/Linguística; Eliza Maria Barbosa – UNESP/Psicolgia da Educação; Rosângela Nogarini Hilário – UNIFESP/Educação; Paula Cristina Bullio – UNICAMP/Linguística; Alessandra Jacqueline Vieira - IFSP – campus São Roque (SP); Lourenço Chacon – UNESP/Fonoaudiologia.

[3] Grupo COLAJE, coord. Aliyah Morgenstern (Université Sorbonne Nouvelle - Paris 3), http://colaje.scicog.fr ; e o grupo de pesquisa da Profa. Dra. Anne Salazar-Orvig (Université Sorbonne Nouvelle - Paris 3), http://www.univ-paris3.fr/salazar-orvig-anne-29869.kjsp .

i) vídeos coletados por curto período de tempo: Mar., 2;5 a 3;2 anos (bilíngue, PB-francês); C., 6;10 a 8;11 anos (bilíngue, PB-espanhol); Me., 2;1 a 2;11 anos (monolíngue, PB); F., 5 a 6 anos (bilíngue, PB-espanhol); An., 1;1 a 2;9 anos (monolíngue, PB);

ii) vídeos coletados por um longo período de tempo: sete crianças monolíngues (G., S., M., Ma., L., O., E.), registradas do nascimento até os 7 anos de idade;

iii) vídeos coletados em meio escolar: S., 6;0 a 10;10 anos (portadora de paralisia cerebral quadriplégica distônica, conta com filmagens realizadas semanalmente, durante 1 hora e com produções escritas); 25 alunos do nono ano do Ensino Fundamental (14-15 anos), de uma escola particular de Sertãozinho, SP (duas aulas/debate registradas em vídeo, além de cerca de 100 produções escritas desses sujeitos, ao longo de 1 ano).

Esse conjunto de dados coletados desde 1996 passou a integrar então o banco de dados do grupo NALingua, em 2008, sendo publicados em teses, anais, capítulos de livros e artigos, pelos membros do referido grupo[4].

Vale dizer que o banco de dados será ampliado ao longo do tempo, com a inserção de novas crianças, não apenas em meio familiar, mas também escolar, respeitando as normas éticas preconizadas na resolução 196/96 do Conselho Nacional de Saúde (CNS), revisto pela resolução 466/2012.

Embora existam outros bancos de dados de fala de crianças no Brasil (UNICAMP, UFPB, PUC-RS, entre outros), o ineditismo deste que está sendo constituído encontra-se: em primeiro lugar, no fato de que, apesar de tudo, são poucos os dados acessíveis aos pesquisadores da área de Aquisição da Linguagem, sobretudo do Português do Brasil (PB) e de crianças bilíngues que têm o PB como uma das línguas; em segundo lugar, porque nenhum outro banco de dados dispõe de um acompanhamento por tão longo tempo com a mesma criança, como é o caso das crianças que foram (e estão sendo) acompanhadas do nascimento até os 7 anos; finalmente, porque os dados existentes não estão sistematizados ou transcritos de forma padronizada, o que inviabiliza qualquer tipo

[4] Parte integrante do grupo NALingua, o GEALin (FCLAr), Grupo de Estudos em Aquisição da Linguagem da UNESP-FCLAr contribuiu para a constituição desse banco de dados. Dele fazem parte alunos de iniciação científica (IC), mestrado, doutorado e pós-doutorado. O objetivo desse grupo é discutir a questão da subjetividade, o posicionamento da criança em relação ao outro no discurso, a partir de dados coletados longitudinal e transversalmente e com base em um referencial teórico que se pauta em Bakhtin e no Círculo (BAKHTIN, 1993, 1997a,1997b, 2006; VOLOSHINOV, 1976).

de trabalho de pesquisa de natureza colaborativa, como os de comparação com dados de fala de crianças de outros países, sobretudo com os EUA ou a Europa.

Diante disso, a proposta de constituição desse novo banco de dados pretende suprir essa carência que existe na área, uma vez que ele poderá beneficiar não apenas os trabalhos que são desenvolvidos pelos pesquisadores e estudantes dos grupos NALingua (e GEALin), mas os de outros pesquisadores do Brasil e de qualquer outro país do mundo que se interessarem em fazer estudos que envolvam o PB, como é o caso da França, país com quem desenvolvemos projetos de colaboração. Isso porque a metodologia de coleta de dados segue padrões rigorosos que vão desde a qualidade do equipamento utilizado até a forma como os dados serão tratados após a sua coleta. Há uma frequência certa para essa coleta, uma postura a ser adotada pelo pesquisador, uma forma de armazenamento dos dados, um programa específico para transcrição e análise de resultados (CHAT/CLAN), enfim, algumas particularidades para que os dados atendam às necessidades atuais da comunidade científica nacional e internacional.

Essa padronização internacional nos permite dialogar com pesquisadores da área do mundo inteiro[5], observar, na mesma criança, todo o seu desenvolvimento linguístico, bem como estabelecer uma relação com o processo inicial de entrada da criança na escrita.

Coleta e organização dos dados: descrição de procedimentos

Para garantir uma boa qualidade dos registros é necessário dispor, igualmente, de um bom material: deve-se utilizar um equipamento de registro cuja qualidade das gravações (áudio e vídeo) nos permita armazená-las por longo período, sem danificá-las.

[5] Vale dizer que já há colaboração com dois grupos de pesquisa na França, mas pelo fato de poucos dados estarem sistematizados, até o momento não foi possível realizar estudos comparativos entre o PB e o Francês. No entanto, as possibilidades de colaboração são bem promissoras. Recebemos este ano (2015) um financiamento da Université Paris 3 – com possibilidade de renovação –, para o desenvolvimento de um projeto sobre humor (Projet RIHA, com Profa. Dra. Aliyah Morgenstern), e que pode ser gasto apenas com deslocamentos (Brasil-França, França-Brasil) e diárias. Além disso, fomos convidados a liderar no Brasil um polo para desenvolvimento de uma pesquisa sobre a aquisição da referência em crianças, em um projeto submetido em setembro à ANR-2015, sob a coordenação geral da Profa. Dra. Anne Salazar-Orvig (Université Paris 3).

A recomendação internacional[6] é que sejam utilizadas câmeras AVCHD com saída para microfone externo, de preferência da marca SONY (câmera e microfone). Considerando a natureza das gravações, é necessário que o microfone (ECM-W1M multi-interace set) capte a conversação de todos os participantes e, considerando que um deles será uma criança possivelmente em movimento, o transmissor wireless é um recurso imprescindível. Deve-se atentar igualmente à capacidade dos cartões de memória, já que para cada hora de filmagem 4GB são necessários, bem como aos computadores que farão as conversões do vídeo e as transcrições, que devem suportar a alta qualidade das imagens gravadas.

Uma vez registrados, os dados devem ser organizados. Os vídeos são gravados, originalmente, no formato MTS ou AVI (em alta definição). Para a transcrição, é necessária a conversão dos vídeos para o formato MOV. Como são, em geral, arquivos grandes (de 2 a 3 GB, que aqui chamamos de *formato G*), é necessário também a compressão dos vídeos (até 800 MB, que aqui chamamos de *formato P*), a fim de facilitar o compartilhamento dos dados. Além disso, deve-se calcular a idade da criança (no formato CHAT, utilizando os comandos do programa CLAN) e em seguida, nomear os vídeos (nome e idade da criança). Esse mesmo procedimento deverá ser feito com os dados que ainda serão coletados.

Os dados são armazenados de diferentes formas - HDs externos de grande capacidade, de 2 ou 3 T, DVDs - havendo pelo menos um backup deles. O grupo de pesquisa dispõe igualmente de um servidor que segue alocado nas dependências da Faculdade de Ciências e Letras/UNESP, campus de Araraquara.

Transcrição dos dados

Uma vez registrados e organizados, os dados são transcritos. No Brasil, alguns pesquisadores inicialmente lançavam mão das normas do projeto NURC[7], que eram

[6] http://www.talkbank.org/info/dv/equipment.html

[7] O projeto NURC (Norma Urbana Oral Culta), foi criado em 1969. Para mais informações, ver CASTILHO e PRETI (1987).

utilizadas integralmente ou de forma adaptada. Dois empecilhos, no entanto, se colocavam frente a essa prática: por um lado, as normas do NURC não contemplavam as especificidades da fala da criança, por outro lado, as adaptações impunham aos pesquisadores restrições no compartilhamento dos dados, já que não havia homogeneidade nas convenções adotadas. Visando superar tais dificuldades, optou-se pela transcrição dos dados pertencentes ao grupo NALingua tendo como base um conjunto de normas reconhecido internacionalmente, o CHAT/CLAN. Como as normas encontram-se em inglês, a fim de facilitar seu acesso aos pesquisadores brasileiros e aos próprios transcritores de nossos grupos, o grupo GEALin desenvolveu uma versão reduzida em português do manual fornecido em inglês e outra com os comandos, ambos publicados em 2012 (DEL RÉ et al, 2012; HILÁRIO et al, 2012).

O CLAN é um programa que está baseado na plataforma CHILDES[8] e, além de fornecer um manual para transcrição (CHAT) inclusive fonética, ele permite alinhar o som e/ou o vídeo à transcrição, estimar um nível geral de linguagem e, sobretudo, medir um eventual atraso de linguagem (MACWHINNEY, 2000). Isto porque o programa CLAN fornece instrumentos para uma análise "automática" de enunciados, palavras e morfemas a partir de uma série de comandos que podem ser criados. Apesar das críticas feitas a respeito dessa contagem (DEL RÉ, HILÁRIO, 2014), as ferramentas do CHILDES dedicadas ao léxico e à morfossintaxe (PARISSE & LE NORMAND, 1998, 2000, 2006), por permitirem uma análise fina dos dados, possibilitam um diagnóstico precoce e preciso e, com isso, o encaminhamento de crianças a profissionais especializados como fonoaudiólogos etc.

Outra vantagem de se utilizar esse programa é a possibilidade de se compartilhar os dados com outros pesquisadores que o adotaram na Europa e nos Estados Unidos, já que eles podem ser disponibilizados em um banco de dados internacional (CHILDES). Nele, é possível encontrar dados de fala de crianças de diversas partes do mundo, mas ainda com poucos dados de crianças brasileiras.

[8] http://childes.psy.cmu.edu

A transcrição feita com o programa CLAN, segundo as normas CHAT, propõe como diferencial, como dissemos, o alinhamento com o vídeo. Assim, quando se clica no enunciado transcrito, aparecem a imagem e o som do enunciado produzido pela criança ou pelo adulto.

Na maior parte dos casos (com exceção feita aos trabalhos que estudam, por exemplo, a linguagem de sinais etc.), o CHAT compreende não apenas a transcrição dos enunciados (a chamada transcrição grafêmica, com convenções da língua escrita), mas de toda a cena enunciativa, podendo incluir linhas adicionais de transcrição fonética, expressões faciais, gestos, entonações, elementos pragmáticos, ação, situação, comentários e outras linhas que se fizerem necessárias, além do alinhamento do vídeo à transcrição (colocação das “bolinhas”), isto é, o recorte do vídeo em enunciados. O que vai determinar os elementos que farão parte da transcrição são os objetivos e as exigências de cada pesquisa. Abaixo, um exemplo de como essas informações aparecem na transcrição:

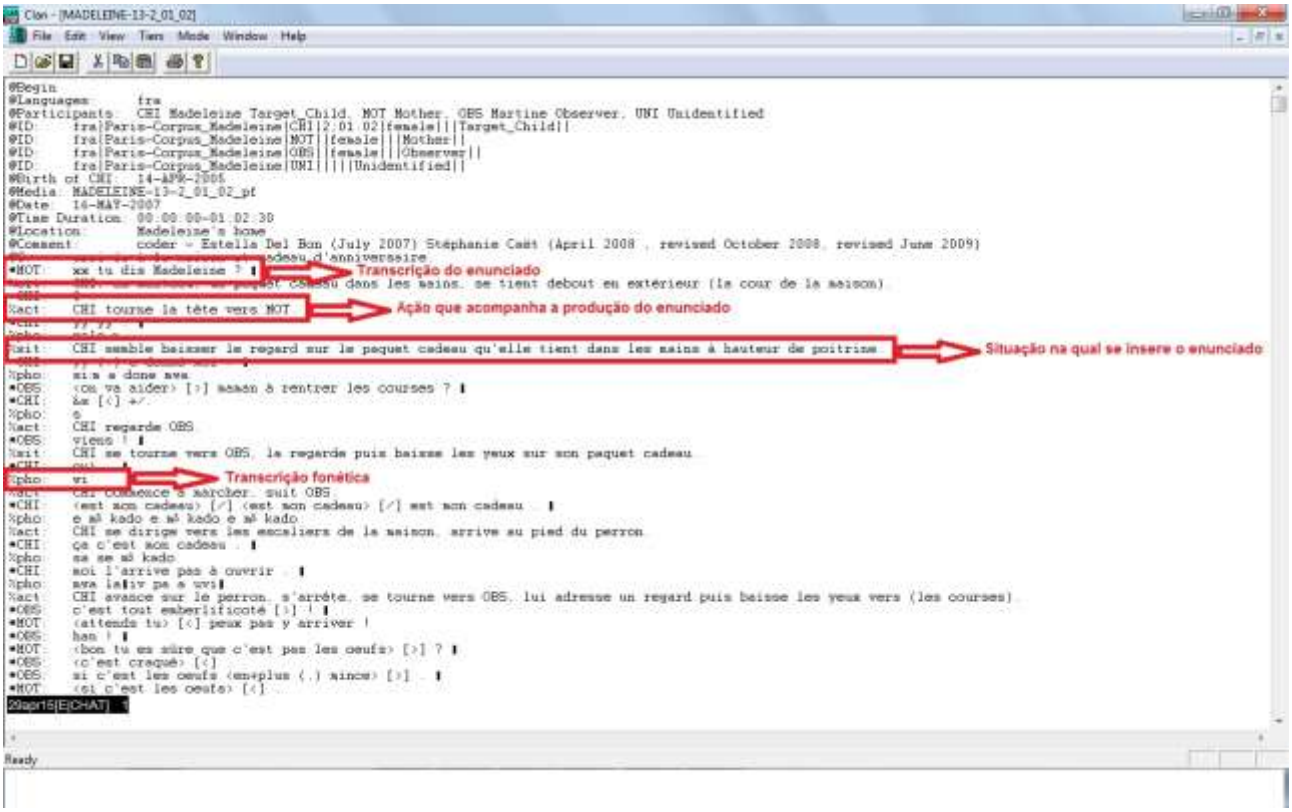

Figura 1: Transcrição CLAN/CHAT

Além disso, todos os aspectos formais devem ser respeitados: o correto preenchimento do cabeçalho com as informações dos participantes da sessão transcrita, o uso adequado dos símbolos do programa etc. Ao final, a transcrição é submetida ao comando CHECK, que identifica possíveis erros (tanto relacionados à sintaxe do programa quanto aos símbolos utilizados). O trabalho de transcrição é concluído com a correção de todos os erros indicados pelo comando CHECK, resultando na mensagem *"Success! No errors found."*.

Como não é possível realizar uma transcrição em tempo real, o ideal é que o pesquisador, sempre que possível, faça anotações durante as gravações ou logo após o término delas: elas podem ser imprescindíveis para solucionar possíveis dúvidas do transcritor.

Vale dizer que, na maioria das vezes, apesar desses elementos, o transcritor ainda necessitará escutar outras vezes as gravações para realizar uma transcrição de qualidade. Às vezes, as crianças falam muito baixo, o som fica distante e nesse caso o vídeo tem um papel fundamental. Em alguns casos o transcritor pode ficar em dúvida em relação ao que a criança teria produzido, ou ainda pode ser que a criança tenha algum tipo de comprometimento de linguagem e, nesse momento, é fundamental realizar uma transcrição fonética. Para uma análise rigorosa da linguagem espontânea das crianças, vale consultar o trabalho de Parisse et Maillart (2004) ou ainda o programa Phon[9], de Yvan Rose (ROSE, 2003).

Quanto mais detalhada é a transcrição, mais tempo ela leva para ser feita e mais recursos financeiros são necessários, portanto, para realizá-la. No caso da transcrição baseada no CHAT, por exemplo, para cada hora de gravação são necessárias cerca de 30 horas de dedicação ao trabalho de transcrição.

Nesse sentido, esse tipo de programa vem de certa forma responder a uma demanda de comprovação dos dados por parte de pesquisadores da área de Ciências Humanas, de um modo geral.

[9] http://childes.psy.cmu.edu/phon/

Revisão da transcrição

A fim de garantir uma maior fidelidade aos dados, além de se poder contar com o vídeo e/ou o som das gravações alinhado a esse tipo de transcrição, os vídeos já transcritos e corrigidos devem ser submetidos a uma revisão. É altamente recomendável que ela seja feita por outra pessoa que não o transcritor. O trabalho de revisão compreende a análise de todos os enunciados transcritos, a fim de confirmar ou questionar o que foi transcrito nas linhas principais (enunciados) e nas linhas adicionais (ação, situação, comentários etc.). Além disso, a revisão compreende também a verificação dos recortes de enunciados. Uma transcrição que contém muitas divergências no momento da revisão pode ser submetida a uma nova revisão, que deverá ser feita por uma terceira pessoa. O objetivo é homogeneizar e confrontar os resultados[10].

Dito isso, é importante lembrar que o transcritor não deve "melhorar" a linguagem da criança no momento da transcrição e deve evitar ao máximo deixar suas marcas, embora, de acordo com Silva (2009), tenhamos que lidar com o "paradoxo do transcritor", isto é, com o fato de ele aparentemente "poder captar tudo para uma escrita oralizada" e, ao mesmo tempo, produzir sempre referências, deixando escapar algo, pois "ao produzir referências no ato de transcrever, o transcritor não consegue apreender o todo" (SILVA, 2009, p. 213).

Criação de uma plataforma para compartilhamento de dados entre os membros do Grupo NALíngua

Com o grande volume de dados que compõem o banco de dados do grupo NALíngua e a necessidade constante de compartilhamento dos mesmo entre os membros do grupo, para fins de organização, transcrição e análise, uma questão importante se colocou: como armazenar e compartilhar de uma maneira mais amigável, porém segura, os dados coletados?

[10] Em francês, esse procedimento recebe o nome de *accord inter-juges.*

Chamamos, aqui, de *compartilhamento de dados* a ação de transferir arquivos digitais, tanto de vídeo quanto de texto (.cha), entre pessoas que têm alguma relação com o grupo de pesquisa NALingua (pesquisadores, pais etc.). O método inicialmente adotado se deu por meio de dispositivos físicos, tais como HD (Hard-Disk), pendrive, ou cartão de memória. A principal característica deles é a capacidade de armazenamento de grandes volumes de dados, na ordem de Gigabytes podendo chegar a Terabytes. O uso de softwares de sincronização online para armazenamento e compartilhamento de arquivos cada vez mais vem ganhando a aceitação por parte de usuários no meio acadêmico. Isso ocorre pelo fato de que neles basta adicionar ou modificar um arquivo em um computador para que, em um intervalo de tempo relativamente curto, outros computadores que também possuam uma instalação do software devidamente configurada se sincronizem, tendo acesso à última versão do arquivo. Alguns softwares possuem uma versão específica para ser instalada em dispositivos móveis como celulares e tablets, o que amplia a possibilidade de acesso aos arquivos. Há ainda em alguns softwares a possibilidade de acessar os arquivos em um servidor online por meio de interface web.

Entre os dispositivos físicos, uma vez que ambos estão em contato direto com a mão do usuário, pode ocorrer deterioração de seus invólucros e perderem compatibilidade física com a interface de conexão. Uma vez que são dispositivos elétricos, estão sujeitos a danos causados por sobretensão. Outro ponto a ser destacado é a sua perda, que pode ocorrer devido ao tamanho físico relativamente pequeno ou pelo seu esquecimento em computadores de acesso público.

Sobre os softwares, cabe destacar dois pontos por meio de questionamentos. Onde estão fisicamente os dados, uma vez que são online e distribuídos globalmente? Quem realmente tem acesso aos dados? Devido às idiossincrasias de cada país, não é possível afirmar que o acesso aos dados de outros usuários em um determinado país seja entendido como algo ilegal, já que a legislação difere de um país para outro.

Como dissemos, o grupo NALingua, a princípio, adotou o uso de dispositivos físicos. Esse modelo ainda vigora devido a sua simplicidade. Em paralelo, foram realizados testes utilizando o plano gratuito do Dropbox, porém viu-se que esse modelo

não atendia às demandas do grupo. Visando oferecer acesso controlado e seguro a usuários de diversos níveis aos dados, o grupo então adotou uma plataforma dedicada.

A plataforma faz uso de um conjunto específico de softwares instalados em um ambiente Cliente-Servidor, com a finalidade de obter gerenciamento e compartilhamento de arquivos. Entre os recursos que uma plataforma dedicada deve oferecer estão a adição, remoção, alteração e compartilhamento baseados em níveis de acesso de usuários. Para sua implementação e manutenção é necessária mão de obra técnica especializada. Suas características, como espaço para armazenamento de arquivos e velocidade do link, estão diretamente relacionadas à infraestrutura do local onde será implantada, assim como à configuração do hardware (servidor) utilizado.

Os dados do grupo NALingua estarão em uma parte do servidor do Programa de Pós-Graduação em Linguística e Língua Portuguesa da UNESP/FCLAr. Para isso, criamos uma plataforma (serviço executado por um técnico em informática) para gerência de conteúdo multimídia. Por meio dela será possível, de forma segura, catalogar e distribuir conteúdo entre membros de grupos de pesquisa. Uma VM (*Virtual Machine* – Máquina Virtual) com duas VCPUs (*Virtual CPU* – CPU Virtual), com 2 GB de memória e 2,4 TB de Disco no data-center da Faculdade de Ciências e Letras de Araraquara – UNESP, hospeda o conjunto de softwares que compõem essa plataforma. O local conta com refrigeração adequada, redundâncias elétrica e do link de dados. A VM é inteiramente open-source, utiliza como sistema operacional o Ubuntu Linux 14.04, rodando o servidor WEB Apache 2.4, que fornece serviço HTTP seguro. É multi-usuário, pois cada usuário tem seu próprio acesso, permitindo o isolamento dos dados e a rastreabilidade dos eventos na plataforma. Para sua utilização são necessários dois papéis de login, administrador da plataforma e usuário. O administrador é responsável por gerenciar contas de usuários e acompanhar eventos na plataforma enquanto o usuário é responsável por adicionar e consultar conteúdo multimídia. A consulta de conteúdo pode ser feita também por visitantes, sendo que para isso o conteúdo deve ser colocado em uma área específica da plataforma. Todas as operações da plataforma, tanto as de adição de conteúdo quanto as de consulta, são protegidas por SSL (Secure Sockets Layer) uma

camada de segurança que impede que terceiros tenham acesso ao conteúdo transmitido. Foi implementada para atender os seguintes requisitos funcionais:

- adicionar, alterar, excluir e consultar conteúdo multimídia;
- operar em modo online, permitindo seu acesso via navegador de internet;
- permitir rastreabilidade tanto de alterações no conteúdo da plataforma, quanto do acesso aos dados.

Em relação a uma plataforma dedicada, pode ocorrer de não ser especificado uma política adequada de cópias de segurança - uma vulnerabilidade que pode acarretar na perda dos dados. Além disso, é possível que ocorram quedas tanto de link de dados quanto de energia elétrica, caso o local de implantação não disponha de redundâncias, deixando a plataforma inacessível.

Por ser uma plataforma multi-usuário, cada usuário tem seu próprio acesso, o que permite isolamento dos dados e a rastreabilidade dos eventos na plataforma. Por meio da figura 2 é possível observar a organização das áreas e dos níveis de acesso da plataforma. Na parte inferior está a infraestrutura, composta pelo hardware e softwares básicos para o seu funcionamento (retângulo cinza). Sobre a infraestrutura há três áreas distintas: a área administrativa (retângulo azul), a do usuário (retângulo lilás) e a do visitante (retângulo laranja). Para seu funcionamento, são necessários dois papeis de login, administrador da plataforma e usuário. O administrador é responsável por gerenciar contas de usuários e acompanhar eventos na plataforma, enquanto o usuário é responsável por adicionar e consultar conteúdo multimídia. Uma vez que o administrador tem acesso à área do usuário, ele pode assumir esse papel executando as tarefas de responsabilidade dele. A consulta de conteúdo pode ser feita por visitantes, sendo que para isso o conteúdo deve ser colocado em uma área específica da plataforma. Tendo em vista que tanto o administrador quanto o usuário tem acesso à área do visitante, ambos podem fazer as mesmas consultas que ele.

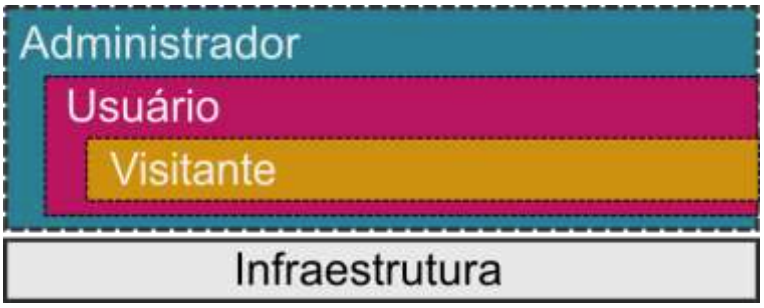

Figura 2: Organização das áreas e níveis de acesso da plataforma

Algumas considerações

Nesse artigo, buscamos demonstrar como a tecnologia tem contribuído para os estudos que envolvem a linguagem, e, mais especificamente, a aquisição da linguagem.

Nosso intuito foi discutir questões metodológicas que interferem diretamente nas pesquisas da área, sobretudo no que se refere à coleta, transcrição e compartilhamento de dados, apresentando soluções práticas para a difícil tarefa de constituição de um banco de dados disponível à comunidade acadêmica.

Muito embora tenham sido elencados pontos de falhas relativos à segurança para cada um dos modos de compartilhamento tratados neste artigo, cabe destacar que, tomando o devido cuidado, ambos funcionam e atingem seu objetivo. O que realmente implica na adoção de um ou de outro é o conhecimento de seus pontos mais ou menos favoráveis, tendo em vista as especificidades dos documentos a serem compartilhados. Há de se levar em conta, ainda, o custo de cada um deles, pois os recursos de um projeto de pesquisa nem sempre são suficientes para que se faça um maior investimento. No entanto, tal custo se dilui se levarmos em consideração os benefícios da constituição de um banco de dados universal e disponível para pesquisadores de diversas áreas, possibilitando o desenvolvimento de pesquisas tanto no Brasil quanto no exterior.

REFERÊNCIAS BIBLIOGRÁFICAS

BAKHTIN, M. *Para uma filosofia do ato*. Trad. Carlos Alberto Faraco e Cristóvão Tezza (tradução não revisada, exclusiva para uso didático e acadêmico) da edição americana *Toward a Philosophy of the Act*. Austin: University of Texas Press, 1993.

_____. O autor e o herói. In: BAKHTIN, M. *Estética da criação verbal*. Trad. Maria Ermantina Galvão G. Pereira. 2 ed. São Paulo: Martins Fontes, 1997a, p. 25-220.

_____. Os gêneros do discurso. In: BAKHTIN, M. *Estética da criação verbal*. Trad. Maria Ermantina Galvão G. Pereira. 2 ed. São Paulo: Martins Fontes, 1997b, p. 261-306.

BAKHTIN, M. (VOLOSHINOV, V. N.). *Marxismo e filosofia da linguagem*. Problemas fundamentais do método sociológico na ciência da linguagem. Trad. Michel Lahud e Yara Frateschi Vieira. São Paulo: Hucitec, 2006.

CASTILHO, A. T. de; PRETI, D., *A linguagem falada culta na cidade de São Paulo*. Projeto N.U.R.C./SP. São Paulo: T. A. Queiroz, Editor/Fapesp, v. II, 1987.

DEL RÉ, A. *Aquisição da linguagem*: uma abordagem psicolinguística. São Paulo: Contexto, 2006.

DEL RÉ, A; HILARIO, R. N. Limites e contribuições do uso do EME para pesquisas de cunho qualitativo na aquisição do PB. *Prolíngua* (João Pessoa), v.8, p.121 - 144, 2014.

DEL RE, A., HILARIO, R. N., MOGNO, A. S. Programa CLAN da base CHILDES: normas de transcrição (CHAT) e comandos básicos In: *Estudos em Aquisição Fonológica*.1 ed.Pelotas : Gráfica e Editora Universitária-UFPel, 2012, v.4, p. 11-30.

HILÁRIO, R. N., DEL RÉ, A. Questões metodológicas e ferramentas de pesquisa nos estudos em Aquisição da Linguagem. *Letras de Hoje*, Porto Alegre, v. 50, n. 1, p. 57-63, jan.-mar, 2015.

HILARIO, R. N., MOGNO, A. S., DEL RÉ, A., VIEIRA, A. J., GRECCO, N., MELLO, I. A. S., BUENO, R. G., FALASCA, P. O CHAT e o CLAN como ferramentas metodológicas nos trabalhos em aquisição da linguagem In: *Na língua do outro*: estudos interdisciplinares em aquisição de linguagens.1 ed.São Paulo : Cultura Acadêmica, 2012, v.1, p. 329-348.

MACWHINNEY, B. *The CHILDES Project*: Tools for Analyzing Talk. 3. ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

PARISSE, C., LE NORMAND, M.T. Traitement automatique de la morphosyntaxe chez le petit enfant. *Glossa* 61, 22-9, 1998.

PARISSE, C., LE NORMAND, M.T. Automatic disambiguation of morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, & Computers* 32, 468-81, 2000.

PARISSE, C., LE NORMAND, M.T. *Une méthode pour évaluer la production du langage spontané chez l'enfant de 2 à 4 ans*, 2006.

PARISSE, C. & MAILLART, C. Les déficits phonologiques des enfants francophones ayant des troubles spécifiques de développement du langage. *Glossa* 89, 34-47, 2004.

ROSE, Y. *ChildPhon*: A database solution for the study of child phonology. 2003

SILVA, C. L. C. *A criança na linguagem*: enunciação e aquisição. Campinas : Pontes, 2009.

SOMMERVILLE, Ian. *Engenharia de Software* 8. ed. Tradução Selma Shin Shimizu Melnikoff; Reginaldo Arakaki; Edilson de Andrade Barbosa. São Paulo: Persson, 2007.

VOLOSHINOV, V. N. Discurso na vida e discurso na arte. Tradução de Cristóvão Tezza para fins didáticos da versão em inglês de VOLOSHINOV, V. N. Discourse in life and discourse in art (concerning sociological poetics). In: _____. *Freudianism. A marxist critique.* Trad. do russo de I. R. Titunik. New York Academic Press, 1976.

Dropbox, Dropbox tour. Disponível em: <https://www.dropbox.com/tour/>. Acesso em 6 de julho de 2016.

Microsoft, Microsoft OneDrive. Disponível em: <https://onedrive.live.com/about>. Acesso em 8 de julho de 2016.

Google, Google Drive - Cloud Storage & File Backup for Photos, Docs & More. Disponível em: <https://www.google.com/intl/en/drive/>. Acesso em 8 de julho de 2016.

SOBRE OS AUTORES:

Alessandra Del Ré: Mestre e doutora em Linguística pela Universidade de São Paulo (USP), realizou parte de seu doutoramento na França, na Université René Descartes (Sorbonne/Paris V), e desenvolveu (2008-2009) uma pesquisa de Pós-Doutorado na Université Paris X/MoDyCo/COLAJE. Desde 2004, é docente do Departamento de Linguística da Faculdade de Ciências e Letras, UNESP, exercendo a função de Professor Doutor. Entre outros trabalhos, organizou, com a Profa Dra. Christelle Dodane (França), um número temático sobre aquisição da linguagem para a revista ALFA (v.54, n.2, 2010) e, em 2014, com as Profas. Dras. Marina Mendonça e Luciane de Paula o livro "A linguagem da criança: um olhar bakhtiniano" (Ed. Contexto), resultado de trabalhos de seu grupo de pesquisa na UNESP de Araraquara - onde leciona - em colaboração com os grupos Slovo e GED. Tem experiência na área de Linguística, com ênfase em Aquisição da Linguagem, atuando principalmente nos seguintes temas: aquisição de língua oral, humor infantil, argumentação, referência, todos eles dentro de uma perspectiva discursiva e dialógica. Desde 2004, desenvolve projetos de pesquisa em colaboração com a Université Paris 3 (Profas Dras. Anne Salazar-Orvig e Aliyah Morgenstern) e Université de Montpellier 3 (Profa. Christelle Dodane). Tem atualmente um projeto com a Profa. Aliyah Morgenstern financiado pela IDEX-Université Paris 3 (Projet RIHA). É líder do Grupo NALingua (CNPq), GEALin (FCLAr), e membro dos GTs de Psicolingüística-ANPOLL (desde 2006), de Argumentação-ANPEPP (desde 2007), dos Grupos COLAJE (França, desde 2008) e DIAREF (França, desde 2009).

Rosângela Nogarini Hilário: Graduada em Pegagogia pela UNESP/FCLAr, doutora em Linguística e Língua Portuguesa pela mesma universidade, na área de Aquisição da Linguagem. Concluiu estágio de doutorado na Université Sorbonne Nouvelle - Paris 3, sob a direção das professoras Anne Salazar Orvig e Aliyah Morgenstern. Atualmente desenvolve uma pesquisa de pós-doutorado sob a supervisão da Profa. Dra. Márcia Romero, na UNIESP/Guarulhos, acerca da problematização e co-construção dos sentidos em sala de aula nas atividades de compreensão textual. Atua nos grupos de pesquisa NALíngua e GEALin, desenvolvendo junto a esses grupos um trabalho de coleta e transcrição de dados, a fim de constituir um banco de dados longitudinais

em português brasileiro a ser disponibilizado na plataforma CHILDES. Tem experiência na área de ensino, tendo atuado como professora tanto no Ensino Fundamental quanto Superior. Tem diversos artigos e capítulos de livro publicados.

Rúbens Antonio Rodrigues: Possui graduação em Tecnologia em Análise e Desenvolvimento de Sistemas pelo Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, (IFSP), Campus de Araraquara (2014) e curso técnico profissionalizante pela Escola Técnica Estadual Doutor Júlio Cardoso - Centro Paula Souza (2006). Atualmente é Técnico em Informática da Universidade Estadual Paulista Júlio de Mesquita Filho. Tem experiência na área de Ciência da Computação, com ênfase em Metodologia e Técnicas da Computação.