RESENHA

O paradoxo entre a transparência dos dados e a privacidade dos informantes na gestão de dados linguísticos

Paloma Batista CARDOSO
Universidade Federal de Sergipe (UFS)

RESUMO

Neste texto, resenhamos o simpósio *Gestão de dados linguísticos* que ocorreu em 21 de julho de 2020 como parte do evento Abralin ao vivo – *Linguists online*, organizado pela Associação Brasileira de Linguística. Nessa atividade, mediada por Raquel Meister Ko. Freitag (Abralin/UFS), os pesquisadores Elisa Battisti (UFRGS), Aluiza Araújo (UECE), Iandra Silva Coelho (IFAM), Marco Antônio Martins (UFSC), Marta Farias Sousa (UFS) e Rodrigo Lopes (UNICAMP) discutiram os aspectos éticos e legais da coleta, armazenamento, tratamento, o uso de ferramentas adequadas, análise e o método da constituição dos bancos de dados linguísticos. As discussões sinalizaram a necessidade de um repositório unificado que poderia ser gerenciado pela Abralin, a fim de possibilitar a comparação entre diferentes variedades do português brasileiro e replicação de estudos, de modo a contribuir para o desenvolvimento dos estudos linguísticos no Brasil e proporcionar uma prática científica transparente, de acordo com o que propõe o movimento *Ciência Aberta*.

ABSTRACT

In this text, we review the symposium *Management of linguistic data* that took place on July 21th 2020, as part of the event Abralin Linguists Online. In this activity, mediated by Raquel Meister Ko. Freitag (Abralin/UFS), professors Elisa Battisti (UFRGS), Aluiza Araújo (UECE), Sandra Silva Coelho (IFAM), Marco Antônio Martins (UFSC), Marta Farias Sousa (UFS) and

OPEN ACCESS

EDITADO POR
Raquel Freitag

AVALIADO POR
Marco Antonio Rocha Martins

DATAS
Recebido: 28/07/2020
Aceito: 17/08/2020
Publicado: 24/08/2020

COMO CITAR
Cardoso, P. B. (2020). O paradoxo entre a transparência dos dados e a privacidade dos informantes na gestão de dados linguísticos. *Revista da Abralin*, v. 19, n. 2, p. 1-9, 2020.

Rodrigo Lopes (UNICAMP) discussed the ethical and legal aspects of collect, storage, treatment, the usage of adequate tools, analysis and the method to constitute linguistic databases. The discussion showed the necessity of a unified repository that could be managed by Abralin, to make it possible to compare different varieties of Brazilian Portuguese and to replicate studies to contribute to the development of linguistic studies in Brazil and provide a transparent scientific practice, according to what the *Open Science* movement proposes.

PALAVRAS-CHAVE
Gestão de dados linguísticos. Armazenamento de dados. Ciência aberta.

KEYWORDS
Linguistic data management. Data storage. Open Science.

Introdução

Neste texto, resenhamos o simpósio *Gestão de dados linguísticos*, que ocorreu em 21 de julho de 2020 como parte do evento Abralin ao vivo – linguists online. Nessa ocasião, os pesquisadores Elisa Battisti (UFRGS), Aluiza Araújo (UECE), Iandra Silva Coelho (IFAM), Marco Antônio Martins (UFSC), Marta Farias Sousa (UFS) e Rodrigo Lopes (UNICAMP) discutiram os caminhos para gestão dos dados utilizados em estudos linguísticos orais e escritos.

O simpósio foi dividido em três momentos: primeiro, a mediadora da mesa virtual, Raquel Meister Ko. Freitag (Abralin/UFS) apresentou a proposta de discussão a partir da qual todas as falas foram estruturadas: os desafios para a gestão de dados linguísticos de acordo com os princípios da ciência aberta. Depois, os seis participantes discorreram sobre coleta de dados, boas práticas para manuseá-los, ferramentas de busca e armazenamento. Por fim, a mediadora direcionou aos seis participantes questionamentos sobre o desenvolvimento de um modelo de gestão unificada que poderia contribuir para o avanço dos estudos linguísticos.

2. O paradoxo entre gestão de dados linguísticos e os princípios da ciência aberta

O simpósio *Gestão de dados linguísticos* teve como objetivo discutir os seguintes questionamentos, elencados pela mediadora em uma fala introduória:

1. Como atender aos princípios de ciência aberta quanto ao armazenamento, reuso e autoria do conjunto de dados linguísticos?

2. Como lidar com a questão entre transparência na ciência e o sigilo dos participantes?

3. Quais as ferramentas mais adequadas para vitalidade dos conjuntos de dados linguísticos?

4. Quais ferramentas permitem melhor armazenamento e um sistema de interface para consulta e pesquisa?

5. Como ficam os grupos minoritários e variedades sub-representadas?

No Brasil, a coleta de dados para fins de pesquisa nas Ciências Humanas e Sociais foi normatizada pela Resolução 510/2016, que considerou que a prática científica deve ser pautada por princípios éticos. Nesse sentido, a relação pesquisador-participante deve respeitar a dignidade, liberdade e autonomia do ser humano, sem lhe causar nenhum prejuízo. No Artigo 3º inciso sétimo, a Resolução 510/2016 reconheceu como um dos princípios éticos das pesquisas em ciências humanas e sociais:

> VII - a garantia da confidencialidade das informações, da privacidade dos participantes e da proteção de sua identidade, inclusive do uso de sua imagem e voz.

No Artigo 5º, que discorre sobre o processo de consentimento e do assentimento livre e esclarecido, a mesma resolução definiu que:

> O processo de comunicação do consentimento e do assentimento livre e esclarecido pode ser realizado por meio de sua expressão oral, escrita, língua de sinais ou de outras formas que se mostrem adequadas, devendo ser consideradas as características individuais, sociais, econômicas e culturais da pessoa ou grupo de pessoas participante da pesquisa e as abordagens metodológicas aplicadas.
> **§1º** O processo de comunicação do consentimento e do assentimento livre e esclarecido deve ocorrer de maneira espontânea, clara e objetiva, e evitar modalidades excessivamente formais, num clima de mútua confiança, assegurando uma comunicação plena e interativa (BRASIL, 2016).

Especificamente nos estudos linguísticos, consentimento e assentimento ocorrem por meio da assinatura, antes da coleta, de um documento que autoriza o uso dos dados para fins de pesquisa. Além da declaração de consentimento e assentimento, na prática sociolinguística, por exemplo, é

comum que os participantes também preencham fichas sociais com informações sobre moradia, escolaridade e renda, dados que, de acordo com os princípios elencados acima devem ser resguardados. Apesar dos inúmeros avanços científicos que ocorreram em todas as áreas, Freitag destacou que movimentos negacionistas têm se tornado comuns, o que configura um momento de crise para a ciência. Por isso, é necessário que a prática científica seja transparente. A fim de suprir essa necessidade, surgiu o movimento *Ciência Aberta*.

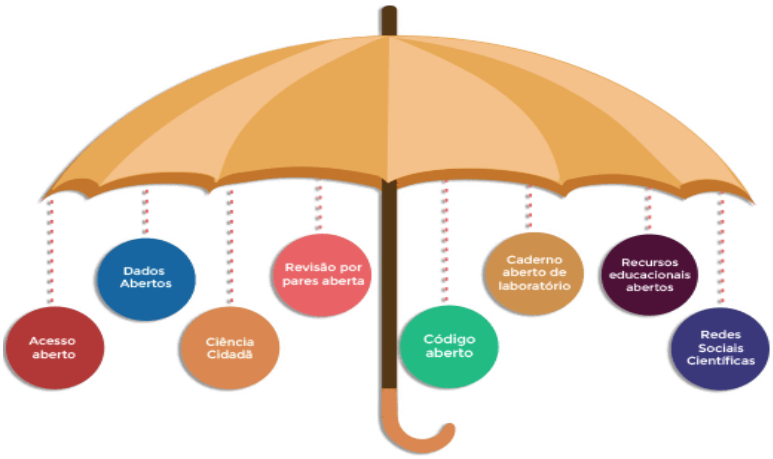

FIGURA 1 - Diretrizes do movimento Ciência Aberta
Fonte: https://url.gratis/v5lhG

Nessa perspectiva, os dados que subsidiam análises devem ser acháveis, acessíveis, interoperáveis e reutilizáveis (*findable*, *accessible*, *interoperable*, *reusable*). Além disso, os métodos utilizados nas pesquisas devem ser registrados no caderno de laboratório para que todos tenham acesso, e o trabalho resultante deve ser disponibilizado em repositórios revisados por pares, de aceso aberto. Apesar dos benefícios trazidos por essas diretrizes, Freitag levantou os seguintes questionamentos: como operar de acordo com o que propõe o movimento *Ciência Aberta* e salvaguardar a identidade dos informantes, conforme estabelecido pela Resolução 510/2016, considerando que os dados analisados nos estudos linguísticos muitas vezes contêm informações pessoais? Como tornar os dados acháveis, acessíveis, interoperáveis e reutilizáveis se, no Brasil, não há um repositório unificado onde todos os pesquisadores que trabalham com descrição linguística possam armazenar seus dados? Diante dessa situação, a Associação Brasileira de Linguística poderia se responsabilizar pela criação e gestão desse repositório? Caso sim, como ele seria gerido? Seria viável estabelecer uma taxa para aqueles que tivessem interesse em acessar os dados? Qual seria o status de autoria desses dados? Foi a partir deles que as discussões expostas na próxima sessão foram realizadas.

3. Gestão de dados linguísticos em bancos de dados brasileiros

Nesta sessão, expomos o que cada um dos seis participantes abordou no simpósio *Gestão de dados linguísticos*. As seis comunicações abordaram o processo de coleta, considerando seus aspectos legais, o tratamento, armazenamento de dados, ferramentas de busca e análise, além dos desafios que o paradoxo entre o que preconizou a Resolução 510/2016 e o movimento *Ciência Aberta* fez surgir.

Depois de elencar os questionamentos expostos na sessão anterior, a mediadora do simpósio concedeu a palavra a Elisa Battisti (UFRGS), que discorreu sobre sua experiência na constituição da amostra *Língua*POA, até aquele momento com 103 entrevistas sociolinguísticas estruturadas de acordo com o modelo Laboviano. A pergunta que direcionou a participação de Battisti foi: "como atender aos princípios de ciência aberta quanto ao armazenamento, reuso e autoria de conjuntos de dados linguísticos?" Inicialmente, a pesquisadora categorizou dados linguísticos como patrimônio que deve ser protegido e preservado. A amostra *Língua*POA não é formada somente pelos arquivos de áudio das entrevistas. Ao aceitarem participar da coleta de dados, os informantes assinaram um termo de consentimento e preencheram um questionário socioeconômico a partir do qual foi possível delinear como a comunidade de fala de Porto Alegre era socialmente organizada naquele momento. Quanto ao tratamento dos áudios e dos questionários, Battisti apontou ações de boas práticas para preservação e compartilhamento: após a coleta, as entrevistas sociolinguísticas foram integralmente transcritas no ELAN, software que permite a transcrição ortográfica de entrevistas e que "facilita enormemente a análise linguística dos dados (por exemplo, para codificação de variantes de variáveis fonéticas)" (OUSHIRO, 2014, p. 46). Nesse processo, os nomes dos informantes foram resguardados e as informações pessoais ocultadas para que suas identidades fossem preservadas.

Diretamente relacionado ao tratamento das entrevistas e à preservação das informações dos participantes está o armazenamento desses dados. De acordo com a Resolução 510/2016, os dados utilizados nas pesquisas devem ser protegidos. Todavia, Battisti explicitou que ainda que a amostra *Língua*POA tenha sido constituída por meio de subsídio do CNPq, ela não foi armazenada institucionalmente. Considerando que, de acordo com os princípios do movimento *Ciência aberta* os dados devem ser acháveis, acessíveis, interoperáveis e reutilizáveis, Battisti apontou a necessidade de sistematização de como é feita coleta de dados, o que pôs em evidência a necessidade de um repositório unificado que poderia ser de responsabilidade da Abralin, a fim de armazenar não somente a amostra de Porto Alegre como das diversas variedades do português brasileiro.

Coleta, boas práticas para tratamento e armazenamento de dados, sistematização metodológica e a necessidade de um repositório unificado também foram pontos presentes na fala de Aluiza Araújo (UECE) que, a partir do questionamento "como lidar com a tensão entre a transparência e o sigilo?" discorreu sobre o *Projeto descrição do português oral culto dos fortalezenses* – PORCUFORT que no momento em que o simpósio *Gestão de dados linguísticos* ocorreu estava em sua segunda etapa, com um total de 87 gravações, 84 inquéritos transcritos e 18 revisados. A primeira etapa durou de 1993 a

1996. A partir desse momento inicial, a referida pesquisadora ilustrou a importância de um repositório para armazenamento de dados com o seguinte fato: uma das entrevistas realizada nos anos 1990 perdeu-se devido à ação do tempo sobre o arquivo.

Segundo Araújo, para a constituição da amostra da segunda etapa do *PORCUFORT* as seguintes práticas foram adotadas: os documentadores foram treinados para a coleta, que foi feita com gravadores eficientes. Além da entrevista, posteriormente transcrita resguardando-se informações pessoais, cada participante, assinou um termo de consentimento livre e esclarecido e respondeu questionários que contribuíram para o delineamento social da comunidade de fala.

Assim como a amostra de Porto Alegre, a de Fortaleza também ficou sobre a responsabilidade da pesquisadora e foi armazenada pessoalmente por ela. No processo de constituição do *PORCUFORT* o treinamento dos documentadores, o método de coleta e a transcrição dos dados foi executada dentro dos parâmetros da primeira fase do projeto. A partir dessa primeira amostra, 65 trabalhos foram publicados. Os pesquisadores ligados ao projeto tiveram acesso aos dados, conforme afirma Araújo, "de mão em mão". Este fato evidenciou pelo menos duas questões que merecem atenção no que tange à gestão de dados: a ausência de um repositório unificado prejudica tanto o acesso à amostra por outros pesquisadores que poderiam replicar estudos ou realizar comparações com outras variedades, quanto o resguardo de informações pessoais, uma vez que não há controle formal sobre quais dados dos informantes são compartilhados. As questões referentes à coleta, tratamento e armazenamento dos bancos de dados elencados expostos por Battisti e Araújo também estiveram presentes na fala de Iandra Silva Coelho (IFAM), estruturada a partir da pergunta: "como atender aos requisitos da ciência aberta quanto à coleta, armazenamento e o reuso de dados linguísticos?".

Coelho discorreu sobre a constituição do banco de dados linguísticos *Reci*, utilizado para pesquisas nas modalidades acadêmica e profissional, sendo que esta última pressupõe a elaboração de uma proposta ou produto educacional. O *Reci* foi constituído de acordo com o protocolo da sociolinguística variacionista e contém dois tipos de amostra: uma oral, restrita para proteção da identidade dos informantes, armazenada pela pesquisadora, e outra escrita, irrestrita, utilizada como base para elaboração de um produto educacional de acesso livre no repositório da instituição.

Considerando que a constituição de amostras de dados linguísticos ocorre em diferentes regiões do Brasil é importante considerar as particularidades de acesso a cada comunidade de fala. Para coleta entre povos originários, especificamente na região Norte, há uma logística distinta: alguns grupos vivem em lugares isolados e para ter contato com eles para fins de pesquisa é necessário obter autorização junto ao ICMBio. Outro aspecto destacado foi o da presença do pesquisador: a comunidade deve aceitar quem fará a documentação linguística que, por sua vez, deve respeitar o que foi proposto pela Resolução 510/2016.

Além dos aspectos referentes à coleta, armazenamento e acesso aos dados do *Reci*, Coelho apontou outro fator que justifica a necessidade de um repositório de armazenamento: há variedades cuja documentação é delicada por conta da dificuldade de acesso. Em um repositório unificado, seria possível que pesquisadores de diferentes lugares tivessem acesso a dados que no momento são

restritos, de modo a impulsionar a realização de estudos linguísticos que incluam as variedades faladas no norte do Brasil.

Após as considerações de Coelho, Marco Antônio Martins (UFSC) discorreu sobre formatos e ferramentas adequadas para armazenamento e consulta de dados linguísticos. O pesquisador apresentou uma plataforma desenvolvida em parceria com a Universität zu Köln, na Alemanha, com objetivo de armazenar e disponibilizar textos de diferentes corpora produzidos no Brasil no curso dos séculos XVIII a XXI e que possam ser analisados em programas para análise estatística como o GoldVarb e o R.

Para Martins, o uso de ferramentas de armazenamento e busca traz para a documentação linguística a necessidade de diálogo com profissionais da tecnologia da informação e do direito, uma vez que os dados linguísticos, que contém informações pessoais, passam a ser tratados como um patrimônio. Os aspectos legais da gestão dos dados linguísticos foram pontuados por Marta Farias Sousa (UFS), que estruturou sua fala a partir do questionamento: "ciência aberta, ética e gestão de dados de fala: como lidar com a tensão entre a transparência e o sigilo?" Farias abordou a Resolução 510/206, pontuando os artigos 3º, inciso 7, e o artigo 5º parágrafo primeiro, expostos na segunda parte desta resenha. Além deles, destacou também o Artigo 9º, que garantiu como direitos dos participantes:

> III – ter sua privacidade respeitada;
> IV – ter garantida a confidencialidade das informações pessoais;
> V – decidir se sua identidade será divulgada e quais são, dentre as informações que forneceu, as que podem ser tratadas de forma pública (BRASIL, 2016).

A pesquisadora reiterou que os artigos citados estabelecem como os dados linguísticos devem ser geridos. Em termos de gestão, dois aspectos foram enfatizados: os dados coletados deveriam ser armazenados em uma plataforma unificada e disponibilizados a partir de tipos de licença, em um modelo que poderia ser semelhante ao da *linguistic data consortium* que especifica, mediante pagamento, o tipo de acesso que cada pesquisador pode obter. O uso de uma plataforma de dados gerida pela Abralin contribuiria para que a prática científica dos estudos linguísticos no Brasil tivesse dados acháveis, acessíveis, interoperáveis e reutilizáveis. Nesse ponto, Farias destacou a necessidade de um acompanhamento jurídico para a gestão dos dados, considerando o tipo de informação presente tanto nas entrevistas quanto em fichas sociais ou em qualquer outro tipo de dado escrito.

O último pesquisador a participar do simpósio foi Rodrigo Lopes (UNICAMP), que também discorreu sobre a importância e necessidade de um repositório que armazenasse múltiplas amostras do português brasileiro. No que tange à coleta e análise de dados linguísticos, Lopes salientou o quanto essa plataforma comum seria importante para o aprimoramento do método, pois ela possibilitaria o compartilhamento do protocolo de coleta e a unificação de boas práticas (ou do caderno aberto do laboratório) como o treinamento dos documentadores, uso de equipamentos adequados e preservação das informações dos participantes. Dessa forma, seria possível contribuir com o aumento da transparência do modo como os estudos linguísticos são executados.

A possibilidade dessa plataforma levanta o debate sobre a autoria dos dados e do seu status na área, conforme pontuou Lopes. A constituição de uma amostra demanda investimento financeiro,

recursos humanos e trabalho em equipe. Nesse sentido, os dados poderiam ser considerados produto do trabalho de todos os pesquisadores que estiveram envolvidos em sua coleta. Por isso, eles deveriam ser armazenados adequadamente e seu acesso deveria ocorrer mediante licenças de uso. O debate sobre coleta, tratamento e armazenamento dos dados envolve também metodologia. O que foi exposto tanto na fala de Lopes quanto na dos demais participantes do simpósio pôs em evidência outra necessidade: a de uma formação de pesquisadores que enfatize métodos de coleta, tratamento, manejo dos dados, de ferramentas computacionais e tecnologias, o que certamente contribuiria para o avanço dos estudos linguísticos no Brasil.

4. Discussão e direcionamentos

Após a fala do último participante do simpósio, a mediadora conduziu o debate sobre os pontos discutidos, em especial sobre a disponibilização de amostras de dados sem que sua autoria se perca, da necessidade de um repositório para armazenamento de dados linguísticos e da importância da discussão sobre métodos de coleta, análise e pesquisa em linguística. Questionados sobre a viabilidade de um possível repositório gerenciado pela Abralin de acesso livre para associados e pago para não associados, todos os participantes concordaram com a proposta, com a justificativa de que somente por meio do armazenamento em conjunto seria possível tornar os dados acháveis, acessíveis, interoperáveis e reutilizáveis. Além disso, o controle pelo repositório, estabelecido por licenças específicas, possibilitaria a preservação de dados dos informantes.

Os pontos discutidos pelos seis participantes e pela mediadora da mesa perpassaram a questão da autoria. Freitag (2017) em uma discussão sobre a origem dos dados salientou que a constituição de uma amostra dificilmente é feita por um único pesquisador. No Brasil, é comum que os bancos de dados sejam constituídos em conjunto, por meio do subsídio de agências de fomento. Mesmo assim, "depois de pronto, um banco de dados, ainda que financiado com o dinheiro público, não é de domínio público: houve um trabalho intelectual por detrás de sua elaboração, desde a concepção até a consecução, que configura autoria" (FREITAG, 2017, p. 9). Por isso, ainda que os dados estivessem em uma plataforma unificada, seu uso deveria ser controlado.

As discussões levantadas no simpósio salientaram a necessidade de repensarmos a formação técnica e metodológica dos pesquisadores, e o modo como enxergamos os dados que subsidiam nossas análises para que nossa prática científica seja transparente. Por fim, a mediadora do simpósio reiterou que para a constituição de um repositório de armazenamento são necessárias outras discussões, que podem ou não estar de acordo com o que foi explanado por ela e pelos seis participantes, uma vez que a realidade de coleta de dados no Brasil é diversa. Há um longo caminho pela frente: necessitamos sistematizar, armazenar e saber como lidar, por meio de ferramentas apropriadas, com um grande volume de dados. Este é um debate amplo e necessário, com o qual todos os participantes do simpósio se comprometeram, uma vez que seria impossível esgotá-lo em uma única ocasião.

REFERÊNCIAS

BRASIL. Resolução nº 510/2016: ética em pesquisa em Ciências Humanas e Sociais. 2016.

DIRETRIZES do movimento ciência aberta. Disponível em: <https://url.gratis/v5IhG>. Acesso em: 27 jul. 2020.

FREITAG, R. M.K. A dadidade (ou dadidão) do dado, *Linguística Rio*, vol.3, n.1, maio de 2017.

GESTÃO de dados linguísticos. Simpósio apresentado por Elisa Battisti, Aluiza Araújo, Iandra Silva Coelho, Marco Antônio Martins, Marta Farias Sousa e Rodrigo Lopes [s.l.,s.n], 2020. 1 vídeo (2h 30min). Publicado pela Associação Brasileira de Linguística. Disponível em: <https://www.youtube.com/watch?v=S7YS57i7ogs>. Acesso em: 25 jul. 2020.

OUSHIRO, L. Transcrição de entrevistas sociolinguísticas com o ELAN. In: FREITAG, R. M. K. *Metodologia de Coleta e Manipulação de Dados em Sociolinguística*. São Paulo: Blucher, p. 46-50, 2014.