16. Lusitanistentag
München 15/09-19/09/2024

Resumos/ Zusammenfassungen

Banco de Dados VARSUL (Variação Linguística na Região Sul do Brasil): contribuições para o estudo do Português Brasileiro

Cláudia Regina Brescancini (PUCRS/CNPq)
Isabel de Oliveira e Silva Monguilhot (UFSC)

O banco de dados Varsul (Variação Linguística na Região Sul do Brasil) reúne aproximadamente 500 gravações em áudio da fala de indivíduos nascidos e residentes em localidades histórica e socioculturalmente relevantes dos estados brasileiros do Paraná, Santa Catarina e Rio Grande do Sul. Armazenadas em arquivos digitais, essas gravações correspondem a entrevistas de experiência pessoal conduzidas à luz da orientação teórico-metodológica da Teoria da Variação (Labov, 2008 [1972]). As primeiras 288 entrevistas, realizadas entre 1989 e 1996, compõem a chamada *amostra base*; o restante do acervo, de aproximadamente 300 entrevistas, foi coletado nos anos seguintes e compõem a *amostra complementar*. A fim de garantir o anonimato dos entrevistados, em atendimento aos preceitos éticos e legais brasileiros, a amostra base está, atualmente, sofrendo o processo de desidentificação e anonimização, o que possibilitará, em futuro próximo, sua disponibilização para a comunidade de pesquisadores interessados. Ao longo de seus vinte e nove anos de existência, as amostras de fala têm subsidiado o desenvolvimento de pesquisas sobre a variação fonético-fonológica, morfossintática, lexical e discursiva das variedades do Português Brasileiro falado na região Sul do Brasil, apresentadas em artigos científicos, dissertações de Mestrado e teses de Doutorado. Diante disso, este trabalho, como foco na amostra base, pretende apresentar (i) os desafios enfrentados no processo de apagamento das informações que possibilitam a associação, direta ou indireta, a um indivíduo e (ii) as contribuições dos dados fornecidos pela amostra base para a descrição da variação linguística no Sul do Brasil.

LABOV, W. Padrões sociolinguísticos. Tradução de Marcos Bagno, Maria Marta Pereira Scherre e Caroline Rodrigues Cardoso. São Paulo: Parábola Editorial, 2008 [1972].

16. Lusitanistentag
München 15/09-19/09/2024

Coleções de dados linguísticos para descrições do Português do Brasil, do Lusitanistentag 2025.

Eliaine de Morais Belford Gomes (UFRR)
Quezia dos Santos Lopes Oliveria (UERJ)

Dentro do campo da Linguística, no que tange à variação no Português Brasileiro, aponta-se a importância da constituição de coleções de dados que facilitem e possam subsidiar as investigações dos diferentes fenômenos em variação na língua.

Como forma de estudar essa variação, propomos a constituição de uma coleção de dados linguísticos para facilitar a sistematização da variação, reconhecendo que os fenômenos de variação e mudança da língua são controlados por fatores internos e externos que atuam de forma ordenada.

Para tanto, temos trabalhado na elaboração de um banco de dados linguísticos na região Norte do Brasil, em particular na cidade de Boa Vista, capital do Estado de Roraima. Nosso trabalho foi dividido em etapas e, atualmente, estamos realizando as entrevistas sociolinguísticas. Nosso intuito é apresentar todas as fases do projeto, desde a preparação teórica até a transcrição e disponibilização do acervo de dados para uso.

Por ser a única capital brasileira no hemisfério norte e, consequentemente, longe das grandes capitais e dos grandes centros urbanos, Boa Vista ainda recebe pouco investimento no estudo das variedades do Português Brasileiro. O corpus aqui em foco seria a primeira descrição da modalidade oral a ser sistematizada na região. Nossa pesquisa, então, busca ampliar a perspectiva de observação, construindo uma amostra de dados linguísticos para o norte do Brasil.

Com respaldo em Weinreich; Labov e Herzog, U.; W.; M. (2006) entre outros autores, seguimos as orientações teórico-metodológicas da Sociolinguística Variacionista Laboviana para a coleta dos dados.

COELHO et al. Para conhecer sociolinguística. São Paulo: Contexto, 2015.
FREITAG, R. M. (Org.) Metodologia de coleta em manipulação de dados em sociolinguística. São Paulo: Blucher, 2014.
LABOV, W. The study of language in its social context. Studium Generale, no 23, p. 30-87, 1970.
LABOV, W. Sociolinguistics patterns. Philadelphia: University of Pennsylvania Press, 1972.
MOLLICA, M. C.; BRAGA, M. L. Introdução à Sociolinguística: o tratamento da variação. São Paulo: Editora Contexto, 2003.
WEINREICH, U.; LABOV, W.; HERZOG, M. Empirical foundations for a theory of language change. In: LEHMANN, W.; MALKIEL, Y. (Orgs.). Directions for historical linguistics. Austin: University of Texas Press, 1968. p. 95-195.

16. Lusitanistentag
München 15/09-19/09/2024

O Projeto C-ORAL-BRASIL e Seus Recursos para o Estudo do Português Brasileiro Falado

Heliana Mello (LEEL/FALE, UFMG)
Tommaso Raso (LEEL/FALE, UFMG)
Bruno Rocha (LEEL/FALE, UFMG)

O projeto C-ORAL-BRASIL, desenvolvido na Universidade Federal de Minas Gerais, tem como objetivo criar recursos para o estudo do português brasileiro falado, seguindo o modelo do projeto irmão europeu C-ORAL-ROM, que compilou corpora de línguas românicas. Os corpora da família C-ORAL incluem transcrições de fala espontânea, arquivos de áudio, alinhamento texto-fala e anotações prosódicas, utilizando para tal, em nossos corpora mais antigos, o software WinPitch e, mais recentemente, o ELAN.

O C-ORAL-BRASIL publicou em 2012 o C-ORAL-BRASIL I (corpus oral informal) e está finalizando a diagramação do C-ORAL-BRASIL II, que abrange corpora de fala formal, de mídia e telefônico. Esses recursos se destacam pelo tamanho e pela qualidade acústica, com gravações realizadas utilizando equipamentos de alta tecnologia. Além disso, o projeto inclui *datasets* padrão ouro (minicorpora), etiquetados informacionalmente, permitindo estudos comparativos entre línguas.

Os corpora são organizados por contextos (privado/público) e tipos de interação (monólogo, diálogo, conversa), com metadados detalhados sobre os falantes. O projeto também adota a Teoria da Língua em Ato (L-AcT), que analisa a fala em unidades prosódicas e pragmáticas, abrangendo unidades informacionais, enunciados e atos de fala. A anotação prosódica é feita manualmente por equipes treinadas, com validação estatística para garantir consistência.

O C-ORAL-BRASIL II introduziu melhorias metodológicas, como scripts automatizados para avaliação da qualidade acústica e revisões ortográficas e prosódicas.

O projeto também expandiu seus recursos com minicorpora de outras línguas, como inglês americano e português angolano, ampliando as possibilidades de estudos translinguísticos.

O projeto C-ORAL-BRASIL, adicionalmente, compilou o corpus COBAI (Corpus Oral de Brasileiros Aprendizes de Inglês), o COLPI (Corpus Oral de Língua Portuguesa Indígena) e está em processo de compilação do BGEST (Corpus Oral Brasileiro de Gestos) e o C-ORAL-ESQ (Corpus Brasileiro de Fala de Pacientes Portadores de Esquizofrenia).

O projeto dispõe de site próprio (www.c-oral-brasil.org) através do qual tem-se acesso a corpora disponibilizados gratuitamente, além da interface DB-COM, através da qual podem-se fazer buscas diversas em materiais selecionados.

Em resumo, o C-ORAL-BRASIL oferece ferramentas valiosas para pesquisas em linguística de corpus, fonética, sociolinguística e pragmática, contribuindo para o entendimento do português brasileiro falado em suas diversas variações.

16. Lusitanistentag
München 15/09-19/09/2024

Procedimentos de coleção de dados para constituição de Corpus ODA português-italiano

Ana Luiza Oliveira de Souza (Universidade de Pisa)
Maria João Marçalo (Universidade de Évora)

Diferentes pesquisas reconhecem os efeitos positivos do bilinguismo tanto no âmbito linguístico (Cummins 1991; Serratrice et al., 2012) quanto no âmbito cognitivo (Bialystok et al, 2008; Bialystok et al 2010), evidenciando que durante a fase de aquisição o falante bilíngue necessita de um esforço cognitivo maior para acessar estruturas mais complexas da língua em áreas da morfossintaxe, como por exemplo, na produção e na interpretação de expressões anafóricas. Na esteira desse pensamento, o presente trabalho objetiva apresentar o processo de coleção de dados de fala realizados para a investigação de Pós-Doutorado, iniciada em julho de 2025, e que tem como escopo discutir acerca da aquisição de pronomes com função sintática de Objeto Direto Anafórico (ODA) de crianças bilíngues português-italiano, nascidas na Itália. O foco dessa pesquisa é a aquisição do Português Brasileiro (PB) em contexto de emigração, entendendo-se que nesse cenário o PB é analisado à luz dos espaços da contemporaneidade, a partir da descrição da sua aquisição como língua minoritária e de herança. Assume-se na investigação que a produção dos pronomes clíticos em italiano requer o conhecimento sólido de tipos morfológicos, sintáticos, semânticos e pragmáticos mais sofisticados. Muitos estudos confirmam que crianças italianas monolíngues com desenvolvimento típico produzem ODA, com opções de uso do posicionamento dos clíticos até os 4 anos, sem realizar desvios de posicionamento padrão (Vender et al 2012, Tedeschi 2006). Tais desvios ocasionais de flexão do clítico ocorrem entre os 3-4 anos, com progressões constantes do uso padrão no decorrer do tempo. Por outro lado, investigações que envolvem a língua italiana em contato com outras línguas (Serratrice et al. 2004, Sorace et al. 2009, Sorace, 2011) têm demonstrado que nas produções linguísticas de crianças bilíngues simultâneas e sequenciais podem ocorrer trocas no emprego dos pronomes clíticos em posição de objeto, devido às influências interlinguísticas. Os pronomes clíticos em italiano são pronomes monossilábicos, do ponto de vista prosódico são mais fracos por não ocorrerem isoladamente. Sintaticamente, ocorrem em posição pré-verbal com verbos finitos, como em *La nonna la bacia* (A avó beija-a), e pós-verbal com verbos não finitos como em *La nonna vuole baciarla* (A avó quer beijá-la). Os procedimentos que fundamentam a realização da proposta de pesquisa visam a aplicar experimentos de produção oral com conversações espontâneas entre a pesquisadora e as crianças durante momentos de jogos, elicitando a produção de clíticos; além desse procedimento, o processo de coleção de dados com característica longitudinal com duração de 16 meses, envolve a aplicação de um teste, denominado T-PEC (Crocetti *et al*, 2021), geralmente usado para diagnóstico de Transtorno do Desenvolvimento da Linguagem em crianças italianas em idade pré-escolar. Uma vez que, como se sabe, a produção dos clíticos também marca os possíveis efeitos da influência interlinguística mais gerais do bilinguismo. A pesquisa está sendo realizada com um número considerável de crianças (6 crianças, ou mais) que estão adquirindo o PB e o italiano simultaneamente, na faixa etária entre os 2 e 6 anos de idade, residentes na Itália. Portanto, a presente comunicação discute a viabilidade desses testes para coleção de dados, seu potencial de reusabilidade para investigação de fenômenos linguísticos, entre os quais OD anafórico, entre outros fenômenos, tais como estrutura lexical, semântica e sintática em dados de fala de crianças bilíngues.

Referências

BIALYSTOK, E., LUK, G., PEETS, K. F. & YANG S. Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism: Language and Cognition*, 13, 2010, p. 525-531.

BIALYSTOK, E., CRAIK, F. I. M., & LUK, G. Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 2008, 859–873.

BYBEE, J. Usage-based grammar and second language acquisition. In P. ROBINSON & N. C. ELLIS (org.), *Handbook of cognitive linguistics and second language acquisition*. Routledge/Taylor & Francis Group. 2008, pp. 216-236.

CROCETTI P., FANCELLI S., COLPIZZI I., SUOZZI A., CROCETTI E., BORGOGNI E. & GAGLIARDI G. (2021) T-PEC: a novel test for the elicited production of clitic pronouns in Italian. Preliminary data, *Clinical Linguistics & Phonetics*, 35:7, 636-662, Disponível em: https://doi.org/10.1080/02699206.2020.1818129 Acesso em: 20 jul. 2025

CUMMINS, J. *Heritage language education*: A literature review. Toronto: Ministry of Education, 1983. Disponível em: http://files.eric.ed.gov/fulltext/ED233588.pdf. Acesso em: 20 jan. 2023

ELLIS, N. C. Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention, and the limited L2 end state. In P. ROBINSON & N. C. ELLIS (org.), *Handbook of cognitive linguistics and second language acquisition*. Routledge/Taylor & Francis Group. 2008. pp. 372–405.

GOLDBERG, A. E. *Constructions*: a construction grammar approach to argument structure. Chicago: The University of Chicago Press, 1995.

GOLDBERG, A. E. *Constructions at work*: the nature of generalization in language. Oxford: Oxford University Press, 2006.

HÖDER, S. Constructing diasystems. Grammatical organization in bilingual groups. *In*: ÅFARLI, T. A.; MÆHLUM, B. (ed.). *The sociolinguistics of grammar*. Amsterdam; Philadelphia: Benjamins, 2014. p. 137-152.

HÖDER, S. Grammar is community specific. Background and basic concepts of Diasystematic Construction Grammar. *In*: *Constructions in Contact:* Constructional perspectives on contact phenomena in Germanic languages. Amsterdam: Benjamins, 2018. p. 37-70.

SERRATRICE L, SORACE A & PAOLI S. Crosslinguistic influence at the syntax-pragmatics interface: subjects and objects in Italian-English bilingual and monolingual acquisition. *Bilingualism Language and Cognition*, 7, p. 183-205. 2004.

SORACE, A., SERRATRICE, L. FILIACI, F. & BALDO, M. Discourse conditions on subject pronoun realization: testing the linguistic intuitions of older bilingual children. *Lingua*, 119, p. 460-477, 2009.

SORACE, A. Cognitive advantages in bilingualism: Is there a “bilingual paradox”? In P. VALORE (org.) *Multilingualism. Language, Power, and Knowledge,* Pisa: Edistudio. 2011, p. 335-358.

TEDESCHI, R. The acquisition of object clitics in Italian: Data from an elicited production task. *Annali Online di Ferrara – Lettere*. Vol. 2 31-42. 2006.

VENDER, M., GUASTI, M. T., GARRAFFA, M., & SORACE, A. *Bilinguismo precoce e Disturbo Specifico del Linguaggio*. Somiglianze e differenze. 2012. Disponível em <http://www.bilinguismoconta. it/wpcontent/uploads/2012/07/studi-aitla_31. pdf > Acesso em 31.01.2024.

16. Lusitanistentag
München 15/09-19/09/2024

Portal digital de estados de coisas em Português e em línguas românicas a variar e ensinar: o percurso de incorporação de coleções de variedades de língua portuguesa

Marcia dos Santos Machado Vieira (Universidade Federal do Rio de Janeiro/CNPq/FAPERJ)
Pedro Giovani Duarte Poppolino (Universidade Federal do Rio de Janeiro/CNPq)

A equipe do projeto Predicar (Formação e expressão de predicados complexos e predicações), da Faculdade de Letras da UFRJ, em parceria com o Grupo de Engenharia do Conhecimento, do Instituto de Computação da UFRJ, tem dedicado sua atenção à concepção, ao planejamento e à estruturação de um repositório digital de dados linguísticos, com ênfase em corpora escritos de variedades de línguas românicas: InCorpora – Portal digital de estados de coisas em Português e em línguas românicas a variar e ensinar. O intuito, com esta proposta, é tratar de elementos do planejamento e do registro de coleções de dados da língua portuguesa em domínio jornalístico (não-)acadêmico brasileiro, pelas quais se iniciou a constituição de amostras. É também abordar aspectos relativos à governança, interoperabilidade e (re)usabilidade. Para tanto, tencionamos apresentar, em nome dessa equipe, os perfis de coleções e o dicionário de metadados que foram até então pensados para a incorporação de amostras de língua portuguesa no Brasil. O espaço virtual InCorpora vem sendo estruturado: (i) para a cooperação científica multiusuário, no intuito de contar com redes de pesquisadores interessados na documentação e análise dos usos de línguas românicas; (ii) para incorporação de novas coleções, além das constituídas por membros das equipes formadas a partir de iniciativas dos projetos Predicar, VariaR (Variação em Línguas Românicas até início de 2025 e atualmente Variações e Variedades em Línguas Românicas) ou HDLinguagens (Humanidades Digitais e Linguagens) e (ii) para a disseminação e popularização científicas, de modo a estimular e propiciar o compartilhamento com a sociedade (indo além da acadêmica) de dados/datsets, de desafios e soluções na estruturação de um repositório para a salvaguarda deles e, especialmente, de conhecimentos (socio)linguísticos oriundos de investigações empíricas de dados e datasets, conforme preconizam recomendações da UNESCO sobre Ciência Aberta [1]. Uma primeira versão do sistema, contendo parte do que pretendemos no portal, vem sendo desenvolvida, na plataforma Dataverse, com o padrão de metadados Dublin Core e com base em princípios FAIR [2]. Espera-se, com esta proposta de comunicação, participar da iniciativa da seção “Coleções de dados linguísticos para descrições do Português do Brasil” com subsídios que propiciem adensar, além do Brasil, o debate da temática transversal e transdisciplinar “dados linguísticos” – que perpassa questões de Ciência e Educação abertas e cidadãs, Humanidades Digitais, Processamento de Linguagem, Salvaguarda de patrimônio imaterial, Acessibilidade, Ética – iniciado por ocasião do evento ABRALIN EM CENA 17 (https://abralin.org/abralin-em-cena-17/ [3], ocorrido, em formato on-line, de 26 de maio a 18 de junho de 2023). O desenvolvimento do Portal InCorpora é uma ação científica que soma esforços de pesquisadores das esferas de Computação, Informação e Linguística, pondo em prática uma interação transdisciplinar prevista pelo Manifesto das Humanidades Digitais [4].

[1] UNESCO. Recomendação da UNESCO sobre Ciência Aberta. Unesdoc, 2022. Disponível em: . Acesso em: 27 abr. 2025.
[2] IBICT. Princípios FAIR. Gov.br, 04 abr. 2022. Disponível em: . Acesso em: 27 abr. 2025. [3] Disponível no site da Associação Brasileira de Linguística/ABRALIN e no canal do YouTube da Abralin Acesso a ambos os links em: 29 de abril de 2025.
[4] DACOS, Marin. Manifesto das Humanidades Digitais. humanidadesdigitais.org, 26 março 2011. Disponível em: . Acesso em: Acesso em: 27 abr. 2025.

16. Lusitanistentag
München 15/09-19/09/2024

Plataforma da Diversidade Linguística Brasileira: plano de dados abertos de variedades do Português

Marcia dos Santos Machado Vieira (Universidade Federal do Rio de Janeiro/CNPq/FAPERJ)
Juliana Bertucci Barbosa (Universidade Federal do Triângulo Mineiro //Universidade Estadual Paulista 'Júlio de Mesquita Filho'/CNPq)
Raquel Meister Ko. Freitag (Universidade Federal de Sergipe/CNPq)

Com o propósito de reunir coleções de dados da língua portuguesa ou informações sobre coleções para acessibilidade a dados da realidade linguística no Brasil e em qualquer lugar do mundo, nasce o projeto Plataforma da Diversidade Linguística Brasileira. Esse projeto, concebido no âmbito do GT de Sociolinguística da ANPOLL/Associação Nacional de Pós-Graduação e Pesquisa em Letras e Linguística, propicia atualmente interação entre as instituições ABRALIN/Associação Brasileira de Linguística e IBICT/Instituto Brasileiro de Informação em Ciência e Tecnologia com a finalidade de viabilizar um ambiente de repositório digital para catálogo de espaços que reúnem dados do Português no Brasil e para depósito e salvaguarda de datasets.

Entendemos que descrições qualificadas da complexidade de usos e variedades da língua portuguesa em jogo num país de dimensão continental como o Brasil não podem prescindir de acesso às inúmeras coleções de dados que pesquisas científicas brasileiras têm produzido. Uma forma de acessar tais coleções é contar com um ambiente que sirva de referência (inter)nacional a qualquer interessado na realidade plurilinguística e multicultural do Brasil. Um espaço dessa ordem serve também para propiciar orientações e/ou para prover salvaguarda e conhecimento estratégicos de patrimônio imaterial a qualquer pessoa que tenha uma coleção de dados que documente usos de uma comunidade e que queira partilhar tais dados de modo a que eles ganhem em (re)usabilidade, conforme preconizam recomendações de Ciência e Educação abertas e cidadãs e movimentos como as seguintes iniciativas: Rede Brasileira de Reprodutibilidade (https://www.reprodutibilidade.org/), GO FAIR BR (https://www.go-fair.org/national-offices/go-fair-brazil/ e http://go-fair-brasil.ibict.br/), HDLinguagens (www.hdlinguagens.org).

Nesta comunicação, queremos expor elementos do projeto brasileiro, particularmente os que dizem respeito a coleções da língua portuguesa no Brasil, esperando partilhar e debater com nossos pares desafios e soluções que já se apresentam e também mobilizar adesões à meta central de construção,

16. Lusitanistentag
München 15/09-19/09/2024

Tarefas para a sociolinguística brasileira em tempos de IA: Plataforma da Diversidade Linguística

Raquel Meister Ko. Freitag (Universidade Federal de Sergipe)

A crescente demanda por modelos de língua em larga escala (LLMs) destaca a urgência de se produzir conjuntos de dados linguísticos que representem adequadamente a diversidade linguística brasileira. Atualmente, os LLMs são treinados com grandes volumes de dados obtidos muitas vezes sem critérios éticos, nem transparência quanto à seleção textual ou controle paramétrico. Embora o aprendizado supervisionado com dados estruturados otimize o desempenho desses modelos, esta tarefa é ainda limitada, por o Brasil carece de bases linguísticas representativas de suas mais de 250 línguas, entre indígenas, de imigração, sinalizadas e as diversas variedades do português.

A constituição de amostras diversificadas e atualizadas, capazes de refletir a realidade multilinguística e multicultural do país, não se resume à pesquisa e descrição linguística. A proposta da Plataforma da Diversidade Linguística está diretamente alinhada aos princípios e objetivos do Plano Brasileiro de Inteligência Artificial, que orienta o desenvolvimento da IA no Brasil: "Desenvolver modelos avançados de linguagem em português, com dados nacionais que abarcam nossa diversidade cultural, social e linguística, para fortalecer a soberania em IA." A proposta de uma Plataforma da Diversidade Linguística demanda a criação de um protocolo unificado de coleta, transcrição, etiquetagem e disponibilização de dados multimodais, associado à construção de uma ontologia linguística e ao mapeamento de línguas e variedades. Esta é uma iniciativa que demanda a mobilização de uma rede metodologicamente e politicamente sensível, visando não apenas o desenvolvimento de LLMs mais justos e inclusivos, mas também a soberania linguística nacional.

A Plataforma da Diversidade Linguística parte do reconhecimento de que, quanto maior a diversidade de línguas e variedades a serem contempladas, mais complexas se tornam as exigências metodológicas para a constituição de conjuntos de dados linguísticos representativos. Por outro lado, a adoção de dados anotados, estruturados e consistentes permite uma redução significativa do volume de dados necessário para o aprendizado de modelos de linguagem, pois esses dados favorecem um treinamento mais eficiente e acurado. A Sociolinguística, com sua tradição de pesquisa, acumula experiências na coleta e análise de dados linguísticos, com base em padrões consolidados. No entanto, historicamente, essas coletas têm sido orientadas por objetivos de pesquisa específicos e isolados, o que gera uma multiplicidade de estratificações socioculturais. Essa fragmentação dificulta a cumulatividade dos dados e compromete a possibilidade de generalizações mais amplas, especialmente no contexto de aplicações em larga escala como o treinamento de modelos de inteligência artificial.

A preparação de uma Plataforma da Diversidade Linguística demanda, inicialmente, uma etapa de mapeamento das amostras linguísticas existentes, articulada à construção de uma ontologia específica que dê conta da realidade multilíngue e socioculturalmente diversa do Brasil. A ontologia é uma estrutura formal e organizada, capaz de representar de forma sistemática os conceitos, relações e atributos envolvidos no processamento, e permite padronizar a anotação dos dados, garantindo consistência entre os diferentes laboratórios e pesquisadores associados

ao projeto. Além disso, a ontologia precisa contemplar a integração de dados multimodais — incluindo camadas de áudio, vídeo, transcrição e trilhas de anotação linguística — e assegurar a inclusão de metadados detalhados sobre cada amostra, como identidade dos participantes, local e contexto de produção da fala, condições da gravação, geolocalização, entre outros fatores relevantes.

Outra ação necessária é o mapeamento nacional das línguas e variedades linguísticas a serem incluídas. Mapeamentos envolvem tanto critérios científicos quanto de viabilidade prática. Embora seja necessário priorizar línguas em risco de extinção, assim como línguas que apresentem diversidade tipológica ou estejam situadas em regiões geograficamente estratégica, os critérios de viabilidade e sustentabilidade costumam ser determinantes, considerando aspectos como o acesso ético às comunidades falantes, a obtenção de consentimento informado e a disponibilidade de financiamento para as atividades de campo.

A padronização metodológica para a criação de um protocolo replicável e padronizado de documentação linguística é o primeiro passo para a constituição de amostras de dados, com contribuições tanto para a pesquisa linguística quanto para o desenvolvimento de tecnologias de inteligência artificial que respeitem e reflitam a diversidade linguística nacional.