CAPÍTULO 22

BRAZILIAN DATASETS FOR TEACHING PORTUGUESE

Marcia dos Santos Machado Vieira
Federal University of Rio de Janeiro
Juliana Bertucci Barbosa
Federal University of Triângulo Mineiro

In Brazil – a plurilingual country with more than two hundred million Portuguese speakers – sociolinguistic diversity is a cultural heritage that, although documented in some linguistic data collections, lacks documentation via nationally defined parameters for the constitution of memory collections. Collections of this heterogeneous heritage can be explored in teaching (inter)actions in the context of mother language and non-mother language, in other discursive actions related to translation, interpretation, subtitling, dubbing, as well as in other areas of knowledge. Aware of the potential of this heritage, six researchers from ANPOLL's Sociolinguistics WG[178] designed a project for a national digital repository for cataloging and/or bringing together existing Brazilian collections and those to be built: the Platform for Brazilian Linguistic Diversity. This chapter aims to provide the reader with information about Portuguese speaking and writing data collections and this project, that is being developed by Brazilian Linguistics Association (ABRALIN), and then to contribute to a work with Portuguese that promotes attention and respect for linguistic and sociocultural diversity, for the ways of life of the population in Brazil.

[178] https://anpoll.org.br/gt/sociolinguistica/ (Access date: October 09, 2021).

It is estimated that there are more than 250 languages in Brazil,[179] if we consider indigenous languages, immigration languages, sign languages (LIBRAS, for example), Afro-Brazilian ones. Thus, individuals and communities in a situation of multilingualism coexist or resist: the number of languages used by an individual can be quite varied; there are those who speak an indigenous language or another language (pomeran language, talian language, for example), in addition to Portuguese. The decree no. 7387 (December 9, 2010), which established the National Inventory of Linguistic Diversity (INDL), officially highlights the urgency of documenting and safeguarding the Brazilian linguistic heterogeneity, as well as its valuation.

Actions aimed at Portuguese as a language of science, education, culture and (national and international) relationship in society require ample illustration of the mapping of its plural linguistic reality and its historical, sociocultural and geopolitical contextualization. Aware of this, six researchers from the WG on Sociolinguistics at ANPOLL (National Association for Postgraduate Studies and Research in Letters and Linguistics) designed the Platform of the Brazilian Linguistic Diversity Project,[180] presented to the Reference Center of the Portuguese Language Museum[181] in August 2021 and, soon after, to ABRALIN (Brazilian Association of Linguistics).[182] The design of this project was exposed in an activity of the 2021 Abralin International Congress (Interab 12) entitled "Collections of data open to society: linguistic and sociocultural memory and potential for (re) use". And, before that, the pillars that support it, science and education that are open to society, were addressed during the UFRJ *Festival do Conhecimento in 2021*, in an activity entitled "Possible futures for sociolinguistic data".[183]

[179] See: https://pib.socioambiental.org/pt/L%c3%adnguas, http://portal.iphan.gov.br/indl, http://prodoclin.museudoindio.gov.br/index.php/documentando-linguas , https://www.museu-goeldi.br/assuntos/colecoes/linguistica. (Access date: October 09, 2021).

[180] MACHADO VIEIRA, WIEDEMER, FREITAG, BARBOSA, PERES e MOLLICA, 2021.

[181] https://www.museudalinguaportuguesa.org.br/mlp/centro-de-referencia/ (Access date: October 09, 2021).

[182] https://www.abralin.org/site/ (Access date: October 09, 2021).

[183] https://www.youtube.com/watch?v=ZrZxsd5QQns. Access: 23 jul. 2021.

BRAZILIAN (SOCIO)LINGUISTIC DATABASES: THE DIGITAL REPOSITORY PLATFORM OF BRAZILIAN LINGUISTIC DIVERSITY (ABRALIN)

In general terms, the *Platform of Brazilian Linguistic Diversity* project consists of organizing a digital repository of speech and writing samples as well as signed language(s) samples, present in the Brazilian territory. This platform must have a powerful technological and archival architecture that is fully capable not only of cataloging and preserving collections of sociolinguistic and historical-cultural data and metadata, but also of enhancing access to diverse information by different audiences and encouraging multiple subjects to cooperate and interoperate in favor of the continuous updating, reformulation, expansion, curation and revitalization of this repository.

On the horizon is the possibility of such a repository becoming a national and international reference, including in Brazilian museum spaces, for documenting samples of the country's linguistic and cultural memory, undertaken by (socio) linguists in interaction with researchers from other areas of science. In addition, it is planned that the platform can be accessed and used by future researchers (from any area of knowledge, such as artificial intelligence, machine translation, diplomacy, historians, among others), by educators, students, and finally, by society in general.

We plan to configure the following objects as cultural heritage: i) the mapping and the national cataloging of linguistic data collections that are throughout Brazil and constituted based on different parameters via an information and metadata system that facilitates automated searches; ii) the construction of a digital tool to serve as a repository for (meta)linguistic (meta)data collections, with indexing to specialized consulting and advisory services in the area of Letters and Linguistics, with products resulting from data searches in this repository for, among other purposes, language teaching, heritage education, field research science, storage infrastructure, patrimonialization, safeguarding and valuing sociolinguistic (meta) data. This digital tool will enable the introduction of new data collections and the registration of demands and feedbacks for the vital improvement of the digital platform and the work with which it is implemented, with accessibility to linguist and non-linguist consultants, Brazilian or not.

As for existing databases that may be the first to integrate this repository or the database catalog intended therein, the situation, in general terms, is as follows: some databases are already available to consultants, others not yet. Among those who are, according to what has been reported so far by managers of Brazilian

databases who responded to the Project "Mapping of (socio)linguistic databases in Brazil" (MACHADO VIEIRA; WIEDEMER; BARBOSA, 2021), there are, by example, these: PORTAL - www.portuguesalagoano.com.br; CORPORAPORT - www.corporaport.letras.ufrj.br; VARPORT - www.varport.letras.ufrj.br; Vertente do Português na Bahia - http://www5.uefs.br/cedohs/; VARSUL - www.varsul.org.br; ALIP - http://www.alip.ibilce.unesp.br; NEIS - https://corpusneis.wixsite.com/home/corpus https://www.facebook.com/NEISUFRJ, HistLing - https://histling.letras.ufrj.br/index.php; VALPB - http://projetovalpb.com.br; Peul - https://peul.letras.ufrj.br/ https://www.facebook.com/peulufrj/; https://www.instagram.com/peul.letras.ufrj/; Gevar - @gevaruftmufu; PorUs - http://porus.sites.uff.br; LinguaPOA - https://www.ufrgs.br/linguapoa/, www.ledoc.com.br; Discurso & Gramática https://discursoegramaticablog.wordpress.com/corpus/.

And among those that are still not accessed via the website (at least so far), are, for example: PortVix (UFES), Falares Sergipanos (UFS), Variação e Mudança no Português do Oeste de Santa Catarina/Variation and Change in Portuguese from the West of Santa Catarina (UFFS), Venda Nova do Imigrante – ES (UFES and PROFLETRAS-IFES), Português falado na região do norte de Minas/Portuguese spoken in the northern region of Minas (Unimontes), Projeto Accomodation/ Projeto Acomodação (Unicamp).

In addition to the sites already mentioned, it is also worth considering others that may favor this broad knowledge of Brazil, of its plurilingual linguistic community, such as: Phonetics-Ponology by Thaís Cristófaro Silva (https://fonologia.org/portugues/); Brazilian Portuguese Historical Dictionary (https://dicionarios.fclar.unesp.br/dhpb/), which has "11,133,739 items and 249,372 forms"; Projeto do Atlas Linguístico do Brasil/Linguistic Atlas of Brazil Project (https://alib.ufba.br/); Projeto Urbanização de Dialetos Rurais/ Urbanization of Rural Dialects Project (1985 data) (http://www.stellabortoni.com.br/), a database with 271 pages containing transcripts of interviews (speech data).[184] These databases can be accessed free of charge by researchers or professors who wish to use such materials for research, lesson planning and production of teaching materials. Unfortunately, this data is still spread across different sites, it would be important to have a platform like the one being planned (Digital Platform of Brazilian Linguistic Diversity) that would gather the samples in a single virtual space.

[184] See also: Portal multimodal/multilíngue para o avanço da ciência aberta nas Humanidades/ Multimodal/multilingual for the advancement of open science in the Humanities, whose preliminary version is at http://cienciaaberta.org, as explained in Berber Sardinha *et al.* (2021).

Databases such as those mentioned above also serve as a reference point to rescue the memory of experiences achieved in field research, storage and transcription of materials, to map what can be improved in the procedures involved in the work of constitution, preservation, curatorship and dissemination of (meta) data banks and to train and manage new cadres of researchers for this work and new (sub)systems of interconnected and interoperable information in a digital environment.

Regarding linguistic preservation, as defended by Freitag (2021), at the Festival do Conhecimento da UFRJ/UFRJ Knowledge Festival (2021), the organization of a collection of linguistic data on speeches and writings (signed or not) of the Brazilian territory is also an action of linguistic preservation in Brazil, with the same principles of conservation and memory as what already occurs in other areas, such as the organization of Seed Banks.[185] Just as a Seed Bank aims to store in order to prevent certain cultures disappear in the event of a worldwide plague or disaster, a repository of linguistic data (such as the *Digital Platform of Brazilian Diversity*) aims to safeguard languages and the memory of a people of a given time, that is, in the case of any culture / language being destroyed, there will always be a taste of its existence

The repository also serves for linguistic education, heritage education, language teaching, even more in a context in which variability, dynamicity, multimodality and multisemioticity are expected as basic properties to the conception of language, including official guidelines of teaching, as the Common National Curriculum Base (BNCC, BRAZIL, 2018), a normative document that defines a set of learning and essential skills that all students must develop throughout the stages and modalities of Basic Education in Brazil. BNCC guides in its text that Portuguese language classes in Brazilian schools should be taught thinking about the linguistic variation that exists in any language community, so that the student will be able to understand that a language is marked by different ways of speaking/writing.

[185] As examples of a Seed Bank, there is, in the interior of a mountain in the Svalbard archipelago, in Norway, the Global Seed Silo, a kind of "safe" for seeds, designed to resist climatic catastrophes and nuclear explosions, where they are deposited , from different plant specimens. Another example is the AL Belo Correia Seed Bank, from the National Museum of Natural History and Science in Lisbon, in Portugal and the EMBRAPA Seed Bank, in Brazil, both of which send seeds to the World Bank (Source: https:// www.embrapa.br/busca-de-noticias/-/noticia/49411189/material-genetico-brasileiro-segue-para-deposito-no-bank-de-sementes-da-noruega).

In addition to teaching Portuguese as a mother language, a collection of data with samples of different Brazilian languages can also be used in teaching and learning Portuguese as a non-native language (or Portuguese for foreigners, Foreign Potuguese Language/FPL), as it will enable the planning of classes and the production of teaching materials that show FPL students the characteristics of Brazilian Portuguese and its wide variety. The recognition of the brazilian linguistic diversity must also be present in the initial and continuing education of (maternal and non-maternal) Portuguese teachers, since, as stated by Barbosa; Freire (2020, p. 654):

> [...] o professor de língua portuguesa deve, entre outras ações: (i) reconhecer e trabalhar, em sala de aula, as características e as peculiaridades do Português Brasileiro; e (ii) tornar seus alunos mais sensíveis à diversidade cultural e linguística, incluindo a presente em sala de aula (FARACO, 2008). Assim, diante dessa problemática, parece-nos urgente refletir sobre a diversidade linguístico-cultural da língua portuguesa no Brasil e no mundo e relacioná-la ao ensino, sobretudo na formação do (futuro) professor.[186]

Finally, access to digital literacy is also one of the digital repository contributions; as it is an ally in cultural enrichment, appropriation of scientific knowledge, as well as an ally in the construction of linguistic respect and autonomy of the subjects involved in an active learning process, as it requires/mobilizes the participation of (scientific and non-scientific) society. And it is worth remembering that: "Os letramentos digitais são resultantes das transformações e avanços tecnológicos, são marcados pelo multiculturalismo, pelo plurilinguismo, pela multisemioticidade"[187] (SANTOS; GROSS; SPALDING, 2017, p. 128).

DISCUSSION

Open and citizen sociolinguistic science and education are paths with the potential to promote (i) the awareness of the relationship between individuals and society(ies), (ii) the sharing and co-construction of knowledge and ways to overcome challenges, (iii) the (socio)linguistic activism and (iv) the structuring of

[186] [...] the Portuguese language teacher must, among other actions: (i) recognize and work, in the classroom, the characteristics and peculiarities of Brazilian Portuguese; and (ii) make their students more sensitive to cultural and linguistic diversity, including that present in the classroom (FARACO, 2008). Thus, given this issue, it seems urgent to reflect on the linguistic-cultural diversity of the Portuguese language in Brazil and in the world and relate it to teaching, especially in the training of the (future) teacher.

[187] Digital literacies are the result of technological changes and advances, they are marked by multiculturalism, plurilingualism, multisemioticity.

a working logistics network of cooperation and partnership. Investing in projects and policies in this sense is a demand in the Open Science era (https://en.unesco.org/science-sustainable-future/open-science):

> In the context of pressing planetary and socio-economic challenges, sustainable and innovative solutions require an efficient, transparent and vibrant scientific effort - not only stemming from the scientific community, but from the whole society. The recent response of the scientific community to the COVID-19 pandemic has demonstrated very well, how open science can accelerate the achievement of scientific solutions for a global challenge.
>
> The Open Science movement has emerged from the scientific community and has rapidly spread across nations, calling for the opening of the gates of knowledge. Investors, entrepreneurs, policy makers and citizens are joining this call.

The use of materials that record sociocultural and linguistic heterogeneity in Brazil via a digital platform can promote social inclusion and justice. By being explored in different actions and interactions that occur in the most diverse teaching and learning environments, the materials on this platform will be able to undo the myth of monolingualism, linguistic prejudices and stereotypes regarding the Brazilian linguistic community and, thus, promote linguistic respect. Access to authentic data from different Brazilian regions and their speech communities will provide knowledge of the ways of living and conceptualizing the world and the relationships of belonging and identity of its subjects. And this can only favor research endeavors aimed at the contrastive analysis of languages and varieties, such as the comparison of Romance languages (in the VariaR Project), centered on data on the reality of use. It is, finally, a way to (re)discover Brazil or spaces of interaction and belonging that are fertilly configured in it and sometimes are either not noticed or are underrepresented. It is necessary to invest in gathering good, authentic and representative data of the Brazilian linguistic community, to teach Portuguese with a view to this community of millions of speakers, naturally without losing sight of the fact that any description of a language is always made based on a sample, in a clipping, which, we want to meet FAIR principles: findability, accessibility, interoperability and reusability standards.

FINAL WORDS

We hope, with the information gathered here and with the *Plataform of Brazilian Linguistic Diversity* project, to change the image of Brazilian Portuguese normally anchored in the linguistic "pseudo-centrality" of the Southeast region as a reference for working with Portuguese. And, with the reconfiguration of this

image, we hope that exemplars until then on the margins or outside the classroom will have their proper place and, thus, expand the possibilities of being, acting and interacting. We understand that a repository of data collections is vital to reference the sustainable and ethical development of (socio)linguistic research and descriptions.

REFERENCES

BARBOSA, Juliana Bertucci; FREIRE, Deolinda de Jesus. A diversidade linguística no ensino de português como língua adicional e língua estrangeira. *In: Estudos Linguísticos* (São Paulo. 1978), 49(2), 2020, p. 651–673. Disponível em: https://webcache.googleusercontent.com/search?q=cache:1PJo5FkeOT0J:https://revistas.gel.org.br/estudos-linguisticos/article/view/2714+&cd=3&hl=pt-BR&ct=clnk&gl=br. Acesso em 10 out. 2021.

BRASIL. Secretaria de Educação Fundamental. *Base Nacional Comum Curricular*. Brasília: MEC/SEF, 2018.

FUTUROS possíveis para dados sociolinguísticos apresentado por Raquel Meister Ko Freitag, Juliana Bertucci Barbosa, Marcos Luiz Wiedemer, Marcia dos Santos Machado Vieira [*s.l.,s.n.*], 2021. 1 vídeo (2h02min). Publicado pelo Festival de Conhecimento da UFRJ (2021) Disponível em: https://www.youtube.com/watch?v=ZrZxsd5QQns. Acesso em: 23 jul. 2021.

MACHADO VIEIRA, Marcia dos Santos; WIEDEMER, Marcos Luiz; FREITAG, Raquel Meister Ko; BARBOSA, Juliana Barbosa; PERES, Edenize Ponzo; MOLLICA, Maria Cecília de Magalhães Mollica. Plataforma da Diversidade Linguística Brasileira. Projeto apresentado à Pró-Reitoria de Pós-Graduação e Pesquisa da UFRJ e à Fundação Universitária José Bonifácio, em razão do Edital BNDES - Chamada Pública para seleção de propostas no âmbito da iniciativa Resgatando a História N. 01/2021, agosto de 2021.

MACHADO VIEIRA, Marcia dos Santos; WIEDEMER, Marcos Luiz; FREITAG, Raquel Meister Ko; BARBOSA, Juliana Barbosa. Mapeamento de Bancos de Dados (Socio)linguísticos no Brasil. Projeto desenvolvido pelo GT de Sociolinguística da ANPOLL e pela Comissão da Área de Sociolinguística da ABRALIN, junho de 2021.

SANTOS, Áurea Maria Brandão; GROSS, Letícia Granado; SPALDING, Marcelo. Conexões entre letramento digital e literatura digital. *Linguagem em foco* - Revista do Programa de Pós-Graduação em Linguística Aplicada da

UECE, v. 9, n. 1, Fortaleza: EdUECE, 2017. Volume Temático: Novas Tecnologias e Ensino de Línguas.