Linguistic Repositories as Asset: Challenges for Sociolinguistic Approach in Brazil

Raquel Meister Ko. Freitag
Vernacular Languages Department
Federal University of Sergipe
São Cristóvão, Sergipe, Brazil
rkofreitag@academico.ufs.br

***Abstract*—This paper provides remarks for a management plan for Brazilian linguistic documentation repositories in order to contribute to their conservation. The depreciation, authorship, sharing, and financing problems are discussed, pointing solutions.**

***Index Terms*—Linguistic repositories management, Brazilian Portuguese, linguistic repository**

I. Introduction

The sociolinguistic approach is characterized by the use of "databases" of authentic speech samples collected through interviews with speakers of a given speech community. The collection of sociolinguistic interviews constitutes a repository of linguistic documentation because the database conception presupposes automated search in a system for storage and organization [1]. Authentic and aligned transcribed linguistic data is an expensive product of interest to those working in data mining, machine learning, and artificial intelligence. In Brazil, this product is a result of larger sociolinguistic projects, with the objectives of:

- Providing resources for the description of Brazilian Portuguese
- Developing and testing linguistic theories
- Training new researchers
- Providing resources for educational programs

This is the case, for example, of NURC [2, 3, 4], PEUL [5, 6], and VARSUL [7, 8]. The intangible assets of the Brazilian sociolinguistic projects were constituted by actions 1 and 4, which constitute the tangible assets. In accounting terms, the difference between intangible and tangible assets is related to depreciation. Tangible assets are those that physically exist; in the case of sociolinguistic projects, their repositories of linguistic documentation are supported. The conservation of these repositories needs to overcome some challenges (depreciation, authorship, sharing, and financing), which are the objective of the discussion in this paper.

II. Challenges

A. Depreciation

In accounting, a tangible asset is depreciable, which means that it loses value over time. While the content of the linguistic corpus is stable, the support for this content is subject to depreciation. In order to beat depreciation, further resources are required. Transferring the audio collection stored on magnetic tapes to digital media, for example, is a procedure that prevents obsolescence (achievement of NURC-Recife [3]), since magnetic tapes have an expiration date and today's devices for this type of media are outdated. Even in digital repositories, routine backup procedures are requested, either in local physical storage media or in cloud storage. There are operational costs involved, both with the storage service and with specialized human resources to carry out this procedure.

B. Authorship

In Brazil, authorship and copyright are regulated by federal law 9610/1998. From a legal point of view, the repository of a sociolinguistic project is assigned as an intellectual property, with copyrights. Thus, the repositories of linguistic documentation from Brazilian sociolinguistic projects are a result of collective construction [9].

From the academic point of view, authorship and contribution are different: a researcher may have contributed to the data collection but may not be considered an author of it [10, 11]. One way of recognizing the types of contribution in science is presented by CRediT taxonomy (Contributor Roles Taxonomy), which names 14 roles that can be assigned to those who contribute to the construction of a scientific product, such as repositories of linguistic documentation. CRediT taxonomy does not attribute authorship, but only formalizes the type of contribution to the scientific product [12, 13]. Also, the CRediT taxonomy specification is more precise than the copyright law.

New linguistic documentation projects have to provide in their design the roles of contributors and copyright. These definitions impact sharing.

C. Sharing

The goal of providing resources for the description of spoken and written Portuguese in Brazil and for educational programs makes the product resulting from the collective undertaking of Brazilian sociolinguistic projects a tangible asset that is not exhausted in itself: sharing is one of the inherent characteristics in the constitution of a linguistic documentation repository [1, 9]. However, although ideally shareable, the circulation of the product takes on barriers associated with

For their support. National Council for Scientific and Technological Development (CNPq), Brazil.

copyrights and ethical aspects, which must be considered in the data management plan.

Since a linguistic repository is an intellectual property product, the legally responsible author (collective work) or the coauthors hold the copyright, which can be Copyright (©) type, which protects the author's exclusive right to take advantage of their product, whether for commercial purposes or not, or Creative Commons (CC), a range of open licenses that encourage reuse and free circulation of authorial products, which involve acknowledging authorship (BY), sharing the product as it is made available (SA), allowing only non-commercial use (NC), or not allowing derivative works from the original (ND).

A linguistic documentation repository may, for example, have a less open license with all rights reserved, or more open licenses that allow reuse but prevent commercial use, or allow unrestricted use as long as the authorship is acknowledged. The data management plan needs to provide the type of license to be assigned to the final product.

D. Funding

Even though speech is free, there are costs involved in making a set of linguistic data systematically organized available in linguistic documentation repositories. To start linguistic documentation, institutional conditions are required: physical space for project allocation and human resources (researchers and assistants). For researchers to be able to elaborate a plan for the documentation and management of the linguistic data, it is necessary to have time allocated for this purpose. In addition, the management of a linguistic documentation project requires specific technical expertise, especially in audiovisual technology, where research assistants available to the project would be the ideal situation (with appropriate pricing in the final product).

After conception, the work team for the constitution of the linguistic samples needs to be trained (which requires the mobilization of a specific structure for this purpose) to develop the specific activities, providing a highly specialized technical service for language documentation. Once the constitution period is over, the data management plan needs to consider the maintenance of the repositories for a long time, or at the project level, which involves annual maintenance costs for as long as the sample remains available, whether the access is more or less open.

Discontinuity of funding accelerates the depreciation of the linguistic documentation repository, which without personal investment becomes obsolete. Loss of tapes, hacking into unprotected servers, and lack of data back-up are all risks arising from the absence of a funding policy for the maintenance of linguistic documentation holdings. It should be emphasized that this is not a problem exclusive to the area of linguistic documentation, but a broader and more systemic problem for all forms of cultural heritage preservation in Brazil.

E. Management plan for Brazilian linguistic documentation repositories

A data management plan is a document that describes the procedures for collecting, processing, organizing, storing, and preserving data, at all stages of a research project. However, not all projects present this plan, not only because it has not been a requirement, but also because some issues still need further discussion.

The Open Science movement for research replicability empowers the repositories of Brazilian sociolinguistic projects as a privileged source for linguistic descriptions. However, the policy of access to the data from these projects is not always explicit to the community. This restriction policy has implications for Open Science requirements, such as those of publications that have as a submission requirement access to the dataset. In Brazilian sociolinguistics, the arguments in favor of a restrictive access policy evoke the waste of time and financial resources in the constitution of the sample, which would generate intellectual property and the right to primacy in the description of linguistic phenomena.

On the one hand, the arguments in favor of an expanded access and sharing policy evoke the nature of public funding of research projects that give rise to products. A linguistic documentation repository is a product subject to all intellectual property laws, and as a product, it should circulate in the community for transparency in research and equity of access to the results, promoting social justice.

The funding argument needs to be relativized because not all costs are covered by project funding, and, even when funding exists (which is not always the case), it is not enough to cover all the steps of the data collection process. Accountability and social justice arising from public funding guarantee the right of access to data, which is not to be confused with total and unrestricted availability; after all, the responsibility for the use and reuse of data is on the authors (responsible researchers, controllers, and organizers).

On the other hand, the starting point of the Open Science movement is transparency and replicability of analysis: does the data actually exist? Will another researcher replicate the same procedures and achieve the same results? Due to this principle, journals have been stimulating the availability of repositories.

Thinking about the sustainability of projects to build linguistic documentation repositories, partnerships with the information technology area, or even companies, could minimize problems of obsolescence and safeguarding of data, by promoting the circulation and automation of analysis through natural language processing algorithms.

These planning actions may help to promote the longevity of the linguistic documentation repositories of Brazilian sociolinguistic research.

REFERENCES

[1] R. M. K. Freitag, et al., “Challenges of Linguistic Data Management and Open Science”, CadLin, vol. 2, no. 1, pp. 01-19, Apr. 2021.

[2] L. A. Silva, “Projeto NURC: histórico,” Linha D’Água, vol. 10, pp.83–90, 1996.

[3] M. Oliveira Jr., “NURC Digital Um protocolo para a digitalização, anotação, arquivamento e disseminação do material do Projeto da Norma Urbana Linguística Culta (NURC),” CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos, vol. 3, n. 2, pp. 149–174, 2016.

[4] M. Oliveira Jr., NURC - 50 anos: 1969-2019. São Paulo, SP: Parábola Editoral, 2019.

[5] M. C. Paiva, and M. M. P. Scherre, “Retrospectiva sociolingüística: contribuições do PEUL,” DELTA: Documentação de Estudos em Linguística Teórica e Aplicada, vol. 15, n. spe, pp 201–232, 1999.

[6] M. C. Paiva, and C. A. Gomes, “Grupo PEUL: passado, presente e futuro de uma agenda de pesquisa,” Cadernos de Estudos Lingüísticos, vol. 58, n. 3, pp.503–519, 2016.

[7] G. Collischonn, and V. O. Monaretto, ”Banco de dados VARSUL: a relevância de suas características e a abrangência de seus resultados, ” Alfa: Revista de Linguística, vol. 56, n.3, pp.835–853, 2012.

[8] L. Bisol, and V. O. Monaretto, “VARSUL e suas origens, uma história sumariada,” Revista virtual de estudos da linguagem–ReVEL, vol. 14, n. 13, pp. vi–xi, 2016.

[9] R. M. K. Freitag, M. A. Martins, and M. A, Tavares, “Bancos de dados sociolinguísticos do português brasileiro e os estudos de terceira onda: potencialidades e limitações,” Alfa: Revista de Linguística, vol. 56, n. 3, pp.917–944, 2012.

[10] A. Brand, et al., ”Beyond authorship: attribution, contribution, collaboration, and credit,” Learned Publishing, vol. 28, n. 2, pp.151–155, 2015.

[11] A. Liz, et al., “Publishing: Credit where credit is due,” Nature News, vol. 508, n. 7496, pp.312, 2014.

[12] A. Liz, A. O’Connell, and Veronique Kiermer, “How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship,” Learned Publishing, vol. 32, n.1, pp.71–74, 2019.

[13] A. Holcombre, “Farewell authors, hello contributors,” Nature, vol. 571, n. 7763, pp. 147–148, 2019.