COLPI

Compilation of an Indigenous Brazilian Portuguese L2 corpus

Gláucia Buratto Rodrigues de Mello°, Heliana Mello*
°UFMG, FAPEMIG *UFMG, CNPq, FAPEMIG

COLPI stands for "Corpus Oral de Língua Portuguesa Indígena" or Indigenous Portuguese Language Oral Corpus. It is a small sized oral corpus which documents Brazilian Portuguese as a second language as spoken by Brazilian Indigenous peoples. In this paper we describe its compilation process and its main characteristics. This corpus represents a first step in the attempt to document and make available data that so far has been scattered and not accessible to researchers. The recordings were carried by an anthropologist in her fieldwork and mostly document narratives, therefore portraying monologic texts. COLPI is part of a larger project aimed at documenting Brazilian Portuguese spontaneous speech, the C-ORAL-BRASIL corpus.

Keywords: Brazilian Portuguese, Indigenous peoples, second language corpus

1. Introduction: COLPI compilation

COLPI is a project dedicated to the compilation of Brazilian Portuguese (BP) spoken as a second language (L2) by Indigenous Brazilian peoples. It is a branch of a larger project the C-ORAL-BRASIL which, on its turn, stands for Spontaneous Brazilian Portuguese spoken corpus as described in Raso & Mello (2012)[1]. The major goal behind COLPI is to document and make available to the larger research community, samples of L2BP as spoken by different ethno-linguistic indigenous groups in Brazil.

The motivation for the creation of COLPI emerged from the availability of previously recorded L2BP files and the opportunity to integrate a portion of them into the C-ORAL-BRASIL corpus. The recordings had been carried in a

[1] www.c-oral-brasil.org

CHIMERA. Romance Corpora and Linguistic Studies 3.2 (2016), 125-132. ISSN 2386-2629

stretch of several years by the anthropologist Gláucia Buratto while doing field work in varied tribal locations in the Brazilian hinterland. The adaptation of these previously recorded data to the C-ORAL-BRASIL specifications were carried under the supervision of Heliana Mello, at the Universidade Federal de Minas Gerais, through a two-year postdoctoral fellowship provided by the Minas Gerais Research Support Foundation (FAPEMIG). The project was undertaken from March 2014 to February 2016.

The parameters for the selection of recordings to integrate COLPI were fundamentally the search for representation of the largest possible variety of ethnicities as well as recording quality. C-ORAL-BRASIL files are all good to high quality wav recordings. COLPI, on the other hand, is comprised by mp3 files recorded with portable recorders with built-in microphones. Therefore, differently from C-ORAL-BRASIL files, COLPI is not adequate for fine phonetic/prosodic analysis. It is therefore geared towards documentation that allows for morphosyntactic, lexical and identity related research. This adds up to other research being carried on in Brazil which aims to study L2BP as spoken by indigenous peoples (Maher 1998; Almeida 2003; Trindade 2009). The goal of having multiethnic representation was achieved as 15 different ethnicities are present in the corpus. As far as content is concerned, COLPI presents an assortment of mythic, cultural and identity narratives that broadly testify to the cultural richness there is to be explored in indigenous narratives.

After the selection of 20 excerpts from the original anthropological recordings, the transcription process was initiated following the parameters established for C-ORAL-BRASIL (Mello *et al*. 2012).

2. Some background: Indigenous languages in Brazil

The 2010 Brazilian population census indicated a total of 890,000 self-declared indigenous individuals in Brazil, from which 817,963 identify themselves as indigenous based on race and color, while 78,900 identify themselves as mixed color but indigenous based on cultural and linguistic identity (IBGE 2010)[2]. From the total indigenous population, about 64% live in indigenous lands or reservations, while about 36% live in urban contexts. It was estimated in the census that there are currently 305 different ethnic groups which speak 274 different languages. It was also established that only about 37% of the total indigenous population speak an indigenous language and 77% speak BP. There is no record of the number of indigenous peoples who speak BP as an L2, nor

[2] http://biblioteca.ibge.gov.br/visualizacao/periodicos/95/cd_2010_indigenas_universo.pdf

the degree of accuracy taken to calculate the percentage of that population who speak Portuguese.

Ribeiro (1995) estimated that at the time of Portuguese arrival at the Brazilian shores in 1500, the Tupi indigenous population was comprised by about one million individuals, Tupi being the largest Brazilian indigenous ethnic trunk grouping several related linguistic varieties. It is not known what the indigenous total population was at that time.

According to Moore (2011) there are two major indigenous linguistic trunks in Brazil: Macro-Ge and Tupi. The largest language families Arwak, Karib, Pano and Tukano. Medium size language families are: Arawa, Katukina, Maku, Nambikwara, Txapakura and Yanomami. Smaller language families are: Bora, Guaikuru and Mura. There are seven identified languages which do not belong to any of the known families: Alkanã, Kanoê, Kwaza, Irantxe, Mynky, Trumai and Ticuna.[3]

The Ethnologue[4], basing their data on other published sources, reports that there are 216 languages presently in Brazil. Of these, there are 201 indigenous languages, of which about 90 are dying.

COLPI is comprised of recordings associated to speakers of fifteen different languages: Aweti, Baniwa, Desana, Fulni-ô, Guarani Nhandewa, Kaigang, Kalapalo, Kamayurá, Kuikuro, Mehinaku, Tariano, Waurá, Xetá and Yawalapiti. These languages are dived into the two major trunks Macro-Ge and Tupi, as well as the language families Arwak, Tukano, Karib and Pano.

As far as indigenous peoples competence in BP, there are no precise estimates; however the prediction is that as schooling progresses in the reservation villages, more and more children will be exposed to bilingual education, therefore increasing competence in BP.

3. COLPI compilation

Except for the Kaxinawá recordings which were obtained in 2015, all other files were recorded between 2005 and 2012. The permissions for use of recorded material for research purposes were mostly issue by tribe's chiefs as a collective document. Some of these were obtained *a posteriori*, as a requirement for the recorded material to integrate the C-ORAL-BRASIL, in compliance with current ethics research guidelines in Brazil. The summary of COLPI recordings is presented in Table 1.

[3] Moore (2011) for a complete list of Brazilian indigenous languages.
[4] http://www.ethnologue.com/country/BR

Table 1: COLPI files

Ethnic group/ number of recordings	**File**	**Recording Date**	**Trunk**	**Family**	**Language**
Aweti: 1	biawemn01	10/11/2005	Tupi	Aweti	Aweti
Baniwa: 1	bibanmn01	18/02/2011		Aruak	Baniwa
Desana: 1	bidesmn01	21/10/2011		Tukano	Desana
Fulniô: 1	bifulmn01	2011	Macro-Ge	Yathê	Fulniô
Guarani: 2	biguamn01	09/01/2011	Tupi	Tupi-Guarani	Guarani
	biguamn02	10/01/2011	Tupi	Tupi-Guarani	Guarani
Kaingang: 2	bikaimn01	06/01/2011	Macro-Ge	Ge	Kaingang
	bikaimn02	11/01/2011	Macro-Ge	Ge	Kaingang
Kalapalo: 2	bikalmn01	15/11/2005		Karib	Kalapalo
	bikalmn02	22/11/2005		Karib	Kalapalo
Kamayurá: 1	bikammn01	03/11/2005	Tupi	Tupi-Guarani	Kamayurá
Kaxinawá: 2	bikaxmn01	28/11/2013		Pano	Kaxinawá
	bikaxmn02	29/11/2013		Pano	Kaxinawá
Kuikuro: 1	bikuimn01	11/11/2005		Karib	Kuikuro
Mehináku: 1	bimehmn01	08/11/2005		Arwak	Mehináku
Tariano: 1	bitarmn01	04/11/2011		Arwak	Tariano
Waurá:2	biwaumn01	01/11/2005		Arwak	Waurá
	biwaumn0	01/11/2005		Arwak	Waurá
Xetá: 1	bixetmn01	21/01/2011	Tupi	Tupi-Guarani	Xetá
Yawalapiti: 1	biyawmn01	05/11/2015		Arwak	Yawalapiti

The specific topics offered by informants, which comprise the COLPI files are the following:

a. Kaxinawá/HuniKwin (provenace: Western Amazonia): oral traditions;
b. Aweti, Kalapalo, Kamayurá, Kuikuro, Mehinaku, Waurá and Yawalapiti (provenance: Northern Mato Grosso): foundational myth also known as Kwarup in the High Xingu reservation area, where the recordings were carried;
c. Baniwa, Desana and Tariano (provenace: Northern Amazonia): traditions and rites of passage;
d. Guarani, Kaingang and Xetá (provence: Northern Paraná): collective interviews about their cultural traditions and current living conditions;
e. Fulni-ô (provenance: Southern Pernambuco): cure rituals and practices.

COLPI comprises 28,319 words and approximately 190 minutes of recording. Some of the files have lengthy periods of silence which have not been edited.

As for informants gender, twenty eight are males, three are females and two informants are not identified as they had serendipitous participation. Their ages are estimated, as indigenous peoples do not time their life by calendar years and neither celebrate birthdates. The estimated ages vary between 25 and 75 years old. Seven informants are traditional healers, seven are school teachers, seven are tribe's leaders, six are tribe's chiefs, one is a craftswoman, one is the director of a village school, one is an agricultural-forest agent and one is a common tribe's person. As far as schooling is concerned, fourteen informants have no schooling, ten have what they labelled some schooling, nine informants have schooling ranging from grade school to incomplete college education.

Given the broad variety as far as informants' profiles are concerned, COLPI portrays a wide variation in BP competence. As expected, older informants are less proficient in BP as they have no schooling and had little if any contact with BP speakers along their lives; on the other hand, younger informants have usually been schooled and might be considered bilingual to some extent. The proximity or distance from indigenous villages to urban centers and country towns also influence the degree of familiarity informants have with BP. Additionally, chiefs and tribe's leaders, as well as teachers, have had more contact with the BP mainstream population as required by their community roles.

4. COLPI: some linguistic features

The overall observation of COLPI files so far reveals several morphosyntactic features commonly found in non-standard BP as spoken in rural areas. A fine grained analysis still needs to be carried in order for a more detailed description

of the data to be achieved. Some of the most striking phenomena observed are illustrated below.

In (1) break of subject verb agreement can be observed as the 3PPL pronoun is followed by a 3PS verb form:

(1) *aí eles começa*
then 3PPL start.3PS
'Then they begin.'

In (2) the lack of SN gender agreement is exemplified:

(2) *Encontra o capivara*
meets the.MASC capybara.FEM
'(He) meets up with the capybara.'

In (3) variable tense marking is indicated where in standard BP there would be simple past marking:

(3) *Os home é corta né cortar derrubar depoi*
the.PL man.SG is cut.PRES uhm cut.INF fall.INF.down after
tirar o galho
take.INF the branch
'The men cut uhm fell down (the tree) and later they took the branch out.'

In (4) a reduced pronominal form *aque* instead of *aquele* can be observed:

(4) *aque que vai levano arco e flexa*
that who is taking bow and arrow
'That who is taking the bow and arrow.'

Some interesting phonetic/phonological phenomena can be observed as exemplified in (5):

(5) a. Roticization /l/ -> /r/: prantá <- plantar 'to plant'
b. Depalatilazion /š/ -> /s/: samá <- chamar 'to call'
c. Vocalization /ƛ/ -> /ij/: oijá <- olhar 'to look'

Many expressive devices such as onomatopoeic expressions and repetitions are used in order to indicate animals, natural and supernatural phenomena and intensity effects, as illustrated in (6) and (7):

(6) *Fufufufufu*
'bird song'

(7) *foi indo foi indo foi indo foi indo*
went going went going went going went going
' (he) kept on going'

5. Final remarks

While COLPI does not pretend to be a representative corpus of L2BP, it neatly captures 15 different ethnolinguistic groups in the vast domain of indigenous language varieties in Brazil. Its main purpose is to start the documentation of BP as spoken as L2 by indigenous Brazilians and make it available to the broad research community. The next steps in the development of the project are a fine grained analysis of the data so that the 20 files can be classified as to degree of proficiency as an additional metadatum. Additionally, the careful identification of morphosyntactic and lexical phenomena needs to be pursued so that a better description of the available material can be achieved. In the prospects are also the full morphosyntactic tagging of the corpus through the PALAVRAS parser (Bick 2000), and last but not least, its entire availability through a digital interface for open research.

Acknowledgments

Gláucia Buratto is grateful for a two year post-doctoral grant (PMPD-2014/2016) awarded by CAPES-FAPEMIG. Heliana Mello is thankful for research grants from CNPq and FAPEMIG and the continued support from the Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) - UFMG team.

References

Almeida, R.H. de (ed.) 2003. *Aldeamento do Carretão segundo os seus herdeiros Tapuios: conversas gravadas em 1980 e 1983*. Brasília: FUNAI/CGDOC.

Bick, E. 2000. *The parsing system PALAVRAS: automatic grammatcal analysis of Portuguese in a constraint grammar framework*. Aarhus: Aarhus University Press.

IBGE 2010. *Censo demográfico 2010: características gerais dos indígenas, resultados do universo*. Published by Instituto Brasileiro de Geografia e Estatística. http://biblioteca.ibge.gov.br/visualizacao/periodicos/95/cd_2010_indigenas_universo.pdf (accessed April 4, 2016).

Maher, T.M. 1998. Sendo índio em português. In I. Signorini (ed.), *Lingua(gem) e identidade: elementos para uma discussão no campo aplicado*, 2nd ed. Campinas: Mercado de Letras/Fapesp, 115-138.

Mello, H., Raso, T., Mittmann, M., Vale, H. & Côrtes, P. 2012. Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação. In T. Raso & H. Mello (eds), *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Ed. UFMG, 125-176.

Moore, D. 2011. Línguas indígenas. In H. Mello, C. V. Altenhofen & T. Raso (eds), *Os contatos linguísticos no Brasil*. Belo Horizonte: Ed. UFMG, 217-240.

Raso, T. & Mello, H. (eds) 2012. *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG.

Ribeiro, D. 1995. *O Povo Brasileiro: A Formação e o Sentido do Brasil*. São Paulo: Cia das Letras.

Trindade, I.E. 2009. O fenômeno da monotongação no português Tapuio. MA diss., Universidade Federal de Goiás.