The VARSUL Database

Odete Pereira da Silva Menon/Edson Domingos Fagundes/Loremi Loregian-Penkal (Paraná)

Abstract

This study introduces the Project that gave origin to one of the most important databases about oral language in Brazil. The Project on Urban Linguistic Variation in the South of Brazil (VARSUL), that started in 1990, initially comprised the three federal universities of the three States of Southern Brazil: Federal University of Santa Catarina (UFSC), Federal University of Paraná (UFPR) and Federal University of Rio Grande do Sul (UFRGS). In 1993, the Project began to also rely on the Pontific Catholic University of Rio Grande do Sul (PUC–RS). The VARSUL Project aims at storing samples of speech realizations by inhabitants of socio-representative urban areas from each of the three states of the South of Brazil, stratified by location, age range, gender and education.

1 Introduction

The VARSUL database results from executing the project of *Urban Linguistic Variation in the South of Brazil*, whose concept was idealized in 1984 by Leda Bisol who had reunited some researchers in Porto Alegre. The Project proposed by that researcher was based on the pioneering project of socio-linguistic survey in Brazil: the Linguistic Census Project of Rio de Janeiro, coordinated by Anthony Julius Naro, and carried out in the late seventies, at the Federal University of Rio de Janeiro (UFRJ), with its first results published in the beginning of the eighties. The Census Project limited data collection to the city of Rio de Janeiro, with interviews performed in different districts, representative of different local communities, especially from the social viewpoint.

Although the data collection model was that of the Census, in order to comprise the ethnic diversity of the region a consensus was established: it would be not enough to make a survey in the capital cities of the three states (Paraná, Santa Catarina and Rio Grande do Sul); it would be required to include some of the ethnic groups representative of the diversified ethnic occupation not only of the South but also of each state, individually. The reason for that was the fact that the project intended to verify whether the Portuguese spoken in the region was different from the other dialects of PB (Brazilian Portuguese) as a consequence of the distinct colonization of these states (this region was nearly depopulated by the time the highest number of slaves came to Brazil). That is why some bilingual informants were interviewed in some places; Portuguese monolingual informants were interviewed in other.

Another difference relatively to the Census Project regards the data transcription system and interview transcripts storage. For Varsul, the system selected allowed for the storage to be done in microcomputers, making data access easier. Data transcription of the South region, because of its own different characteristics, required a system of idiosyncrasy indication, especially as to pronunciation, hence the selection of a three-line transcription system: the first line records the actual syntax of the informants' speech; the second line records pauses and variable phonetic aspects and the third line records the morphosyntatic classification and the remarking of prosodic aspects such as speech emphasis and speed.

2 A brief history of the VARSUL Project

On August 19, 1982 in Porto Alegre, after invitation of the Applied Linguistics Center coordinated by Margot Levi Mattoso from the Instituto de Letras of the Federal University of Rio Grande do Sul (UFRGS) under the direction of Nora Then Thielen, professors representing the three universities of Southern Brazil: Federal University of Santa Catarina (UFSC), Federal University of Paraná (UFPR) and Federal University of Rio Grande do Sul (UFRGS) – reunited for a meeting. They then had the objective of discussing resources and means to make the studies dynamic in the areas of Linguistic Geography, Bilinguism and Linguistic Variation. Thus, a regional project formed by three work groups: (*i*) Linguistic and Ethnographic Atlas; (*ii*) Bilinguism; (*iii*) Linguistic Variation – was born. This project should reach post-graduate studies so that a descriptive material of the spoken language of Southern Brazil was produced.

The group of Linguistic Variation initially discussed the proposal of Leda Bisol to organize a linguistic database of the Southern Region of Brazil comprehending the states of Paraná, Santa Catarina and Rio Grande do Sul. The idea was to develop research along the same lines of the Linguistic Variation Census of the state of Rio de Janeiro, currently known as PEUL (*Programa de Estudos sobre o Uso da Língua* – Program of Studies on Language Use), and coordinated by Anthony Naro. In 1983, the second meeting of the large group took place in Florianópolis.

In October 1984, at the time of a Master examination board at UFRGS, Leda Bisol reunited professors Carlos Alberto Faraco from UFPR, Solange Lira from UFSC and Gisele Machline de Oliveira e Silva, of the Census Project staff from Rio de Janeiro, in Porto Alegre. The objective of that meeting was to discuss the preparation of the Linguistic Variation Project. The meeting also reunited professors Odete Pereira da Silva Menon (UFPR), Clarice Knies (UFRGS) and students Laura Quednau (Scientific Introduction) and Cristina Schmitt (Master course).

This meeting gave origin to the project that, under the name Urban Linguistic Variation in Southern Brazil (VARSUL), was headquartered at four universities: UFRGS, UFSC, UFPR and PUCRS (included in 1993) guided by the general objectives of offering: (*i*) resources for the description of the spoken Portuguese in the country; (*ii*) conditions for testing and development of linguistic theories; (*iii*) conditions for formation of new researchers and (*iv*) resources for educational programs, promoting knowledge about and respect to linguistic varieties.

When preparing this Project, relying on the consultancy of Giselle Machline de Oliveira e Silva (UFRJ), the intent was to take into account the minimum number of five informants by cell, which would result a too large sample, with costs that would not be covered by fostering agencies. As a result, four cities of each state selected to be part of it – ethnically or culturally expressive[1] – would be represented by a set of 24 interviews, 96 by state, 288 interviews in total, each one with approximate duration of sixty minutes.

Data collection started in the state of Rio Grande do Sul, in 1988, and in the other states in 1990, completing the basic sample in 1996, the year when the Data Base was officially inaugurated at the *I South Cone Linguistic Variation Meeting* between September 2 and 4 in Porto Alegre at the UFRGS.

[1] **Rio Grande do Sul**: Porto Alegre, Flores da Cunhas, Panambi and São Borja. **Santa Catarina**: Florianópolis, Blumenau, Chapecó and Lages. **Paraná**: Curitiba, Londrina, Irati and Pato Branco.

Throughout its implementation and consolidation, the general coordination[2] was subsequently undertaken by researchers of the different universities headquartering VARSUL. Initially, it was coordinated by Carlos Alberto Faraco (UFPR). After him came Solange Lira (UFSC), Cecília Inês Erthal (UFPR), Paulino Vandresen (UFSC), Ana Maria Sthal Zilles (UFRGS) and Odete Pereira da Silva Menon (UFPR), Maria Tasca (PUCRS) and, presently, Izete Lehmkuhl Coelho (UFSC).

The final preparation of the VARSUL Project was in 1985, but approval and release of the first resources by FINEP – *Financiadora de Estudos e Projetos* (Funding Agency of Studies and Projects) only took place in 1989, under the management of Cecília Inês Erthal and, from 1993 onwards, of Paulino Vandresen, which lasted for all the period of sampling collection. With the purpose of integrating the activities of different teams, the annual meetings continued to be held, coinciding with the period of accomplishment of the *Círculo de Estudos Lingüísticos do Sul* (CELSUL – Circle of Linguistic Studies of the South).

Regarding data survey methodology, the Labovian line was followed, inspiring the transcription of interviews in the work accomplished by the team of the Censo Project from Rio de Janeiro, although now configured to be stored in microcomputers. Data were transcribed in three lines: the orthographic transcription in the first line, the indication of variations in the second line which allowed for an immediate electronic relation to be established between orthography of a form, usually uniform, and its diverse realizations; in the third line the morphosyntatic classification of the issues as well as some records of speech style.

Once transcribed, data were electronically stored. The interviews originally recorded in cassette tapes are being gradually transferred to CDs – a stage already concluded at PUCRS and still ongoing in the other centers.

It is important to remark that the VARSUL Database has been expanding with the addition of new samples. To the basic sampling formed by informants divided into three education levels, gender and age range, others have been added: a new age range (15–24 years) and one more education level (graduates), the latter only for capital cities. Also, VARSUL has been turning into a privileged reference in the formation of new researchers, opening its doors to undergraduate students (with scholarships for scientific initiation), master and doctoral students.

3 The Region included in the sampling

The VARSUL Project was conceived with the objective of installing a linguistic database in a short term to allow, further ahead, the description of the urban linguistic variation of the Southern Region of Brazil and its local dialects.

The implantation of this database was initially made with data from the capital cities and one inland city of each state. At a second stage, these data were supplemented to cover urban areas more representative of the regions which, under the historic social and cultural viewpoints, stood out and were relevant in each of the states of the Southern Region of Brazil. Thus, for each state four cities were selected, representing the groups proven relevant to its occupation process. The selection criteria of these groups and of the municipalities representing them are presented as follows.

[2] Besides relying on a general coordinator, each headquartered relies on its local coordinator who is expected, among other duties, to take care of the maintenance and expansion of the Data Base. These local agencies have been managed by successive coordinations: Leda Bisol, Clarice Bohn Knies, Ana Maria Sthal Zilles and Valéria Neto Monaretto at the UFRGS; Paulino Vandresen, Izete Lehmkuhl Coelho and Edair Maria Görski, at the UFSC; Cecília Inês Erthal, Iara Bemquerer Costa and Odete Pereira da Silva Menon, at the UFPR and Maria Tasca and Leda Bisol, at the PUCRS.

3.1 Rio Grande do Sul

The state, with a tradition strongly based on the peasant life of the farms, inhabited by indigenous people, Castilians and troopers who arrived from other parts of the country, was only attached to the Portuguese Crown around 1750, under the name of Província de São Pedro.

It brings in its core the history of three important immigrations: the Azoreans, who arrived around 1750, a little less than a thousand people in groups of sixty couples – the so-called "couples of number" – distributed along the areas aimed at settlements, along river margins, gave origin to the cities of Porto Alegre (early Porto dos Casais), Taquari, Osório among others. Together with the Portuguese people who arrived later, they devoted themselves to cattle-breeding activities, extending through the pampas (grassland), where part of the state culture comes from.

The German immigration, represented in the sampling by Panambi, started in 1824 with 43 immigrants, interrupted in 1830 and re-started in 1844, reaching a large contingent. They received colonial land lots along the fertile margins of River dos Sinos and River do Caí and kept spreading along Serra Geral (local mountain range). They were devoted to cattle-breeding and small scale handicraft, the origin of the current industry, remarkably the shoe industry. They are accounted for erecting cities such as São Leopoldo, Novo Hamburgo, Taquara, Panambi (of late immigration) among others.

Flores da Cunha represents the Italian immigration, started with three Milanese families in 1875. They were assigned with the challenges of the plateau and its difficult access ways which, in a way, directed farms to vine growing. They were awarded with the colonies of Conde d'Eu – nowadays the city of Garibaldi, and Dona Isabel – modernly the city of Bento Gonçalves. They grew wheat and vines, and continued inaugurating towns such as the ones mentioned above and others like Veranópolis, Farroupilha and Caxias do Sul.

In addition to these, there are border populations with Argentina and Uruguay, of Spanish language, forming an important part of the population of the State of Rio Grande do Sul, with specific socio-cultural and economic traditions. Among the cities of this region, Livramento, Itaqui, Uruguaiana and São Borja stand out, and the latter was chosen to be part in the sampling.

3.2 Santa Catarina

The sampling of the State of Santa Catarina intended to represent Portuguese as spoken by the descendants of the most expressive ethnic groups of the state: Azoreans (Florianópolis), Italians (Chapecó), Germans (Blumenau) and the highlanders (Lages).

The Azoreans arrived at the shore of Santa Catarina in the period between 1748 and 1756. Basically, they occupied São Francisco do Sul, Nossa Senhora do Desterro (presently Florianópolis) and Santo Antônio dos Anjos de Laguna.

The march of colonization and settlement of the territory of Santa Catarina was re-started in the middle of the Nineteenth Century and it is characterized mainly by European immigration flows. The first colony of Santa Catarina occupied by Germans was Colony of São Pedro de Alcântara. Two decades later, the large German immigration flow took place in this state with the colonization of the middle valley of River Itajaí and the North-east lands of the state, near São Francisco do Sul. As a private enterprise by Herman Blumenau, the colony Blumenau was born in 1850, in the Middle Itajaí-Açu River. Following, the colonies of Dona Francisca (1851), Itajaí-Brusque (1860) and Ibirama (1899) were inaugurated.

In the hydrographic basin of the Itajaí River the first Italians to arrive to Santa Catarina were settled, in the colony "Blumenau", along the margins of the affluent rivers of Itajaí-Açu River; in the colony "Brusque", along the margins of Itajaí-Mirim River and its affluents and then it was transferred from the valley of Itajaí-Mirim to the valley of Tijucas, settling along the Braço River and its affluents. At a later stage, within the basin of Itajaí River, along the margins of Luís Alves River a colony with this same name was born. In a later flow a large amount of Italian settlers came to the valley of Tubarão and then, little by little, went on to other valleys such as Urussunga, Mãe Luiza and eventually Araranguá. Finally, the middle and far West of Santa Catarina were occupied by Italian and German immigrants from Rio Grande do Sul, represented in the sampling by the city of Chapecó.

Lages was also colonized by Italians at a later period after the initial occupation, when it was founded by people of the administrative division of São Vicente/São Paulo (together with indigenous groups from the Jesuitical Missions) in the Eighteenth Century and the following occupation from the *gauchos* established because of the troops' track open from Viamão (RS) towards Sorocaba (SP).

3.3 Paraná

The State of Paraná presents a extremely diversified linguistic panorama: the main reason for such diversity is in the various origins of the population of the state, formed from different flows of population groups: the Portuguese colonizer from the first centuries, the European and Asian immigrants (19th and 20th centuries) and the Brazilian migrants of the last decades, especially from Minas Gerais, São Paulo and Rio Grande do Sul. Therefore, several modalities of Portuguese are spoken in the state.

Since there is no systematic survey of such varieties, the project has tried to comprehend this ethnic occupation by considering the following areas: The North of the state was populated by people from the states of Minas Gerais and São Paulo during the period of coffee plantation expansion (in 1930) and the city of Londrina is representative of this territory occupation. In the South-west and West, the language brought by the colonists from the states of Rio Grande do Sul and Santa Catarina, responsible for the agricultural occupation of that area of the state, will be represented by the city of Pato Branco.

In the region concentrating the immigration of Slavian peoples (Russian, Polish and Ucranian) and still remaining partially bilingual, the city of Prudentópolis (Ucranian) would be more representative but, for not being a urban center sufficiently stratified for the sampling, the city of Irati (Polish and Ucranian), a larger town (although the majority of the population is not bilingual there) was chosen instead. In the Center-south, also called Old Paraná – the area making the state unique the most from the linguistic standpoint, Curitiba – the capital city of the state – was selected.

4 Sampling constitution

In an initial definition of profiles required in the sampling of urban population, the following social characteristics considered significant in previous sociolinguistic researches were taken into account: gender (male and female); age range (25–45 and over 50) and education (elementary, middle and high school).

It was determined that every municipality should be represented in the sampling by a group of 24 interviews, corresponding to 12 profiles (2 genders × 3 education levels × 2 age ranges), each represented by two interviewees. After the definition of these profiles, informants in different districts with a considerable permanent population were sought.

In addition, the speakers had necessarily to fulfill the following pre-requisites: (*i*) to speak only in Portuguese (requirement for interviewees in the capital cities, but not in bilingual areas); (*ii*) to have lived in the city for at least 2/3 of their lives; (*iii*) not to have lived outside the region for more than a year during the period of native language acquisition (2 to 12 years); (*iv*) not to cause awkwardness to other inhabitants of the region.

Neither illiterate nor graduate people were included in the initial stage for the fact that they are the target population for studies on dialect (illiterate) and formal regional urban norms (graduates).

The age range below 20 was not considered for not presenting the linguistic consistency required to the objectives of the initial study (frequent idiosyncrasies). The decision of not including this age range also took into account the sampling size. Since the database had to include 4 municipalities of each state, the number of informants for each municipality was reduced, considering only the social characteristics proven significant in previous studies.

5 Data collection

Data collection was carried out in two stages. With the help of local leaders (priests, teachers, community leaders) the speaker with a compatible profile in each of the districts was sought; the interviewer introduced himself/herself to them usually accompanied by a person from the community, an acquaintance of the interviewee, and identified himself/herself as a university student, asking for cooperation for an academic paper required about how he/she lived, what he/she thought, how he/she entertained, what the actual inhabitant of the city believed.

Once the consent to the interview was obtained, a recording session of about 5 to 15 minutes took place, justified to the speaker as a test of his/her voice in the recorder although, in reality, worked as data confirmation and record of social characteristics of the informant, such as: *home history* (to confirm family and school history), *family history* (to confirm occupational history), *school history* (to confirm education level of interviewee and his/her acquaintances), *occupation history* (to verify social level, jobs, salary level, ambitions etc.), *reading habits*, TV, radio, sports, parties, "hobbies" (to verify socio-cultural characteristics) and the *contact* with speakers of other languages and dialects of Portuguese (to verify any possibility of foreign interference).

Once the first interview was recorded, the second one was set up, lasting up to one hour, performed whenever possible at the interviewee's home so that he felt as little tense as possible despite the normal interference of the recorder (observer's paradox).

The second interview should always be based on the information obtained in the first interview so that the interviewer could prepare a script of subjects to allow the interviewee to feel comfortable and speak for the most part of the time, producing a linguistically varied speech as to vocabulary, structures, verb tenses and modes, pronouns etc. This interview also followed the blind technique, that is, for the speaker the interview had the objective of collecting information about how the actual inhabitant of the city lived, and not about their linguistic marks.

In many cases, the interview relied on the participation of a third person, labeled **intervenient**, who could be either an interviewer's company or a family member of the interviewee. The interview style, despite the informality degree achieved, can also be characterized as semi-distended, although many informants were able to relax during the interview.

6 The transcription model

The VARSUL Project has been designed to allow for analyses to be made at several linguistic levels: phonetic-phonologic, morphologic and syntactic. To accomplish this, following the

Linguistic Census Project model, a three-line transcription system was adopted, according to the example below:

```
       |-----------------------------------------------------|-------|
0121 1 | instalado,  e  em  funcionamento.  *(risos f)       |       |
     2 |  0       1                      3                   |       |
     3 |     i       c  p        s                           |       |
       |-----------------------------------------------------|-------|
```

The first line corresponds to the transcription of the informant speech; however, it is not the faithful reproduction of the speech because the transcription is recorded according to the official orthography of the Brazilian Portuguese, except for small exceptions. For example, the verbal concordance of *tu*, personal pronoun for the second-person singular, should be marked in the first line and then, in the second line, the absence of the concordance –**s** should be indicated.

```
       |-----------------------------------------------------|-------|
0005 1 | que  esta  errado? *Não  pode  porque  tu  estás    |       |
     2 |   e        #r               e    r  e      00   0   |       |
     3 |  c    v      j       a    v     c      n     v      |       |
       |-----------------------------------------------------|-------|
       |-----------------------------------------------------|-------|
0006 1 | ensinando  tudo.(est)*Então  tu  tens  que  fazer   |       |
     2 |                                         --       0  |       |
     3 |     i       n           a    n    v      c    i     |       |
       |-----------------------------------------------------|-------|
```

The second line (*i*) brings information about sound fading and inclusion; (*ii*) indications of pauses (short and long) corresponding to punctuation or not; (*iii*) phonetic specifications of phonemes performed in different manners throughout the various cities for example[3]; pronunciation of /t/ before /i/ and in front of /e/ in an atonic syllable whose performance can be either apical dental or palatalized (with enhancement of the e).

The third line contains the morphological classification of all words.

6.1 What do the three lines stand for?

Engesis[4] developed a program to edit the interviews that, along with the Interpretator, allows for the execution of searches by combining items of two or more lines. This software was developed in a Windows environment and it is aimed at facilitating data search.

For example, it is possible to search all occurrences of **nós** (personal pronoun for the first-person plural) without the corresponding nominal concordance (*nós ia*, instead of *nós iamos*): the word *nós* should be typed in the search box, and the number of words before and after the occurrence; since usually the verb follows the pronoun immediately, the search can be limited to 3 or 4 words after the item.

[3] The notation *#r* in transcript 0005 above is meant to indicate the alveolar trill (as opposed to the tap).

[4] Software copyright owner.

```
        |----------------------------------------------------|-------|
0325 1  | contava,  né?  *Dormindo  eu falava  o  que  nós   |       |
     2  |         1    1    ur                               |       |
     3  |  v       av,p        i      n  ++ v    n  n    n   |       |
        |----------------------------------------------------|-------|
        |----------------------------------------------------|-------|
0326 1  | íamos  fazer.  *Como  às  vezes  dessas  façanhas  |       |
     2  |  000      01                                  00+ |       |
     3  |  v       i       c    pd    s      pn        s     |       |
        |----------------------------------------------------|-------|
```

In a cross-search it is possible to raise the productivity of preposition **a** in the corpus: an *a* should be typed with space before and after it (the number of words before and after it can be indicated) and, in the third line the information about the morphologic class – preposition (p) – is given, which will prevent article *a* to be selected (d).

```
        |----------------------------------------------------|-------|
0025 1  | estrada  de  chão  [pra]- pra  ir  à  escola. *Com |       |
     2  | i          e       5    :        +   +  +i     3   00+|       |
     3  |   s      p     s      p      p    i  pd  s       p |       |
        |----------------------------------------------------|-------|
        |----------------------------------------------------|-------|
0026 1  | uma  sacola  de  pano,  (ruído) na  época  a  gente|       |
     2  |                         1          +            e|       |
     3  |  d     s      p   s           pd    s    d#  s,n |       |
        |----------------------------------------------------|-------|
```

The program executes the search and lists all occurrences found in that interview, which can also be printed.

Data can be used in several analyses after the social factors recorded in each interview (gender, age range, education level, location) as shown by some studies already available in the site *http://www.pucrs.br/fale/pos/varsul/index.php*.

7 Database Use Standards

The VARSUL Linguistic Database is open for consultation and data supply to professors and students of the universities participating in the project (UFRGS, UFPR, UFSC and PUCRS) as well as to the researchers and graduate students connected to both national and international teaching and research institutions, according to the rules found in the VARSUL site, mentioned below.

8 Present Status of the VARSUL Project and Database

On the VARSUL web page (*http://www.pucrs.br/fale/pos/varsul/index.php*) it is possible to find the present structure of the Project's coordination groups as established in the last meeting on October 29, 2008, during the VIII CELSUL – *Centro de Estudos Lingüísticos do Sul* (Center for Linguistic Studies of the South), held at UFRGS in the capital city of Porto Alegre.

On the project's web page is also possible to find further information about the VARSUL Project and the Database as well as about studies produced such as master and doctoral dissertations, graduation monographs and scientific initiation. The areas of research concentration are: Phonological Variation, Morphological Variation, Syntactic Variation, Sociofunctionalist Studies, Sociolinguistics, Syntactic-Discursive-Pragmatic and Text Linguistics.

Presently a database expansion is being accomplished with the inclusion of a younger age range between 15 and 25 years (12 interviews) and graduates, 15 years of education and up (8 interviews). This stage is nearly concluded in the State of Rio Grande do Sul, and at the digitalization and transcription stage in the States of Paraná and Santa Catarina. Jointly with the expansion through addition of the graduate range the group in Rio Grande do Sul also decided to make a re-contact sampling of NURC-POA[5] with the purpose of working not only with research in apparent time – an important objective of the Project, but to make short duration real time research (NURC carried out in 1970) as well. This will allow to recover NURC data and to prepare panel studies (linguistic variation in the community and in the individual, according to LABOV, 1994).

Another objective of the group of professors of the VARSUL Project is the construction of a digital sampling to be made available *on-line* to the researchers interested. Under the coordination of Professor Izete Lehmkuhl Coelho, a funding project for the production of such sampling (recently approved) will allow to expand and make access easier to a part of the Database.

References

Bisol, Leda/Menon, Odete Pereira da Silva/Tasca, Maria (2008): "VARSUL, um banco de dados". In: Votre, Sebastião/Roncarati, Cláudia (eds.): *Antony Julius Naro e a lingüística no Brasil: uma homenagem acadêmica*. Rio de Janeiro, 7 Letras: 50–58.

Knies, Clarice/Costa, Iara B. (1995): *Manual do usuário do banco de dados lingüísticos VARSUL*. Projeto VARSUL. Curitiba, Florianópolis, Porto Alegre. Unpublished Handbook for VARSUL Database users.

Labov, William (1994): *Principles of linguistics change: internal factors*. Oxford: Blackwell.

Projeto VARSUL. Available at http://www.pucrs.br/fale/pos/varsul/index.php (accessed November 2008).

[5] Formal Urban Norm Project in Brazil: recordings made in the cities of Recife, Salvador, Rio de Janeiro, São Paulo and Porto Alegre.