Mapping Paulistano Portuguese: the SP2010 project

Ronald Beline MENDES, Livia OUSHIRO
University of Sao Paulo
Av. Prof. Luciano Gualberto, 403 - Sala 16, Cidade Universitária, 05508-010 - São Paulo - SP
rbeline@usp.br, livia.oushiro@usp.br

Abstract

This paper reports on the objectives, methods, and results from the project SP-2010 (Mendes, 2011), currently under the execution by the *Grupo de Estudos e Pesquisa em Sociolinguística* (GESOL-USP). Its main objectives are (i) to build a contemporary and representative sample of São Paulo Portuguese; (ii) to develop studies of sociolinguistic variation in the city, an understudied speech community (Mendes, 2009; Rodrigues, 2009); and (iii) to make the corpus of recordings and transcripts available online for a wider group of researchers. The first phase of the project aims at collecting 60 sociolinguistic interviews with speakers stratified by sex/gender, age, and level of education by 2013. In view of the highly heterogeneous sociodemographic make-up of the city of São Paulo, fieldworkers also observe distinctions in informants' social class, family generation in the city, and area of residence. Interview recordings follow Variationist Sociolinguistics premises (Labov, 1984, 2006; Tagliamonte, 2006) and data transcription norms are designed as to facilitate automatic data handling in softwares such as R.

Keywords: spoken corpus; Paulistano Portuguese; variationist sociolinguistics; data collection; transcripts.

1. Introduction

Although São Paulo Portuguese has already been documented and analyzed through broad and significant research projects such as Projeto NURC-SP (Castilho & Preti, 1986, 1987; Preti & Urbano, 1988, 1990) and Projeto Para a História do Português Paulista (Castilho 2007), most works within these projects aim at analyzing "Brazilian Portuguese," either in contrast with European Portuguese (e.g., studies on parametric variation), or in relation to its internal processes of change (e.g., studies on grammaticalization).

Among the very few works about Paulistano Portuguese in its social context, Rodrigues (1987) analyzed variable subject-verb agreement (e.g., *nós vamos* vs. *nós vai* 'we go') in the speech of 40 (semi-)illiterate speakers in two favelas, and Coelho (2006) analyzed the variable use of 1PP pronouns (*nós* vs. *a gente* 'we') in the speech of 24 speakers living in a working class community. Yet, to date, little is known about the linguistic production and perception of many other (supposedly) typical Paulistano variants (e.g., the realization of coda /r/ as a tap in words such as *porta* 'door,' the diphthongization of nasal /e/ in words such as *fazenda* 'farm') and other variants in the city, as well as their social distribution and evaluation in the speech community at large.

This may be due to the difficulties of building a representative speech corpus of a heterogeneous and multicultural city with more than 11 million people, highly diverse in terms of their geographical origin, socioeconomic class, and cultural background. According to a recent survey by the Instituto de Pesquisa Econômica Aplicada (IPEA, 2011), 46% of the adult working population (between 30 and 60 years old) living the the São Paulo Metropolitan Area were not born in the state of São Paulo (see Figure 1). Although the survey does not refer exclusively to the city itself, it gives an idea of the intense presence of non-native inhabitants in this region. One can consider that the number of non-Paulistanos living in the city may be even greater, since the 54% of Paulistas include all people born in the state of São Paulo and not only the capital city.

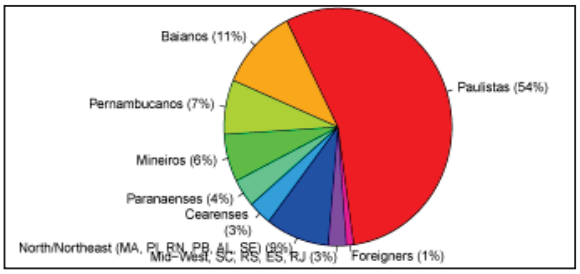

Figure 1: Adult population living in the São Paulo Metropolitan Area. Source: IPEA 2011

This fact raises a number of questions: which social parameters are most relevant for linguistic differentiation and stratification and how to reach speakers of varied social networks? How to gather detailed ethnographic information from each informant (Poplack, 1989), acknowledging a persistent point made by the "third-wave" of sociolinguistic studies (Eckert, 2005) on the importance of observing individuals' social practices? Which methodologies are best for handling a large amount of spoken linguistic data?

In this paper, we report on the objectives, methods, and results from the Project SP-2010 (Mendes, 2011), currently under execution by the *Grupo de Estudos e Pesquisa em Sociolinguística da USP* (GESOL-USP),[1] which aims at: (i) building a contemporary and representative sample of Paulistano Portuguese; (ii) fostering the development of sociolinguistic studies in the city; and (iii) making the corpus of recordings and transcripts available online for a wider group of researchers.

[1] http://linguistica.fflch.usp.br/gesol.

Heliana Mello, Massimo Pettorino, Tommaso Raso (edited by), *Proceedings of the VIIth GSCP International Conference : Speech and Corpora*
ISBN 978-88-6655-351-9 (online)

2. Methods and Results

In 2009-2010, GESOL-USP collected 82 sociolinguistic interviews with residents of the city of São Paulo, native or not to the city, of both sexes and different sexual orientations, from 15 to 89 years of age, with different levels of education, of varied socioeconomic statuses, living in 59 different neighborhoods in the city. In view of São Paulo's great sociodemographic complexity, these exploratory recordings had the objective of defining the most relevant social variables for the sociolinguistic description of Paulistano Portuguese; elaborating an interview schedule; developing best practices in approaching possible informants; identifying possible technical and methodological problems that may occur during the recordings (e.g. avoiding noise, making the informant comfortable) and coming up with solutions for them; and elaborating criteria for transcribing the interviews.

From this experience, we observed that certain sociolinguistic profiles are hard to locate – for instance, younger native Paulistanos who have not concluded at least high school, especially women living in more central areas, or people over 70 who were actually born in the city, especially in more suburban areas. In addition, in spite of our initial aim of locating prototypical speakers from certain neighborhoods (e.g. Mooca, Bexiga, Pinheiros), geographic and socioeconomic mobility seems to be characteristic of the city and its inhabitants, many of whom prefer not to settle in a single place for life. Further, a technical but not to be ignored challenge is the presence of noise (traffic, constructions, people), even in residential areas of the city. The methods designed for this project try to address some of these issues.

In the present phase, to be concluded by 2013, the social parameters for constituting the sample are sex/gender (men and women), three age groups (20-34 y.o.; 35-59 y.o.; 60+ y.o.), and two levels of education (up to high school; college). As our focus is on the *social meaning of variation* (Chambers, 1995), these variables have been chosen primarily because of their potential to shed light on the relationship between variable linguistic uses and social identities, as well as to enable cross-comparisons with other linguistic corpora of Brazilian Portuguese – e.g. VARSUL (Bisol *et al.*, s/d), VALPB (Hora, 2004), PEUL (Paiva & Scherre, 1999), ALIP (Gonçalves, 2003).

Sex/Gender and Age have been broadly analyzed in sociolinguistic studies and have been shown to be correlated with variables whose variants are differently evaluated in terms of prestige: a number of works have have observed that the prestigious forms in the community tend to be employed by women (Chambers, 1995; Labov, 2001; Cheshire, 2004), and that unprestigious forms tend to be avoided by speakers in the intermediary age group, who mostly suffer pressures of the linguistic market (Bourdieu, 1991; Labov, 2001). Correlations with Age can also point to possible changes in progress in the linguistic system through *apparent time analyses* (Labov, 2001). The three age groups are mostly based on their relative position in the job market, but also take into account each group's general lifestyles in a big city. The younger speakers, those between 20 and 34 years old, comprise young adults who tend to be relatively less stable than people in the other two age groups; in São Paulo, it is not rare to find people up to 34 years old who are not married, who do not own their own place, who go to college or who lead life more similarly to people in their early 20s. The group aged between 35 and 59 years old, in turn, is intended to comprise people more fully inserted in the job market and relatively more stable. Finally, the group over 60 years old refers to people in or close to retirement.

Level of education is also directly associated with stigmatization and prestige. The general hypothesis is that more educated speakers will tend to avoid unprestigious forms in the community, or otherwise that the forms they employ will be considered more "correct." In Brazilian sociolinguistic studies, the division between "educated" and "uneducated" speakers is normally taken as an index of socioeconomic status (Rodrigues, 2009: 151). This situation seems to be changing in São Paulo as well as in many other urban centers through extensive public policies of improved access to primary, secondary, and higher education (for instance, *Progressão Continuada* in the state of São Paulo and *ProUni* in a national scope); the division between only two levels of education is a consequence of these changes. However, general increase in average levels of education is not always followed by a direct ascension in individual socioeconomic status, which means that the equation between level of education and social class should not be overestimated. We suggest that level of education should be treated as constitutive of speakers' social class, but not as its substitute.

The combination of these social parameters yields 12 sociolinguistic profiles (e.g. men between 20-34 y.o. without a college degree), each of which is to be filled by 5 speakers, in a total of 60 sociolinguistic interviews. Each of these 5 speakers per cell should reside in a different zone of the city (North, South, East, West, Central), and each cell should contain at least one speaker of three city areas (Downtown, Extended Central Area, Suburbs), as a way to ensure a broad coverage of the city. The speakers' place of residence is defined as the place where he/she has lived for the most part in the past 10 years.

In a second stage, we will focus on social class, a social factor generally overlooked in Brazilian sociolinguistic studies due to lack of reliable criteria for categorizing speakers in different socioeconomic groups (Rodrigues, 2009; Mendes, 2011). In the city of São Paulo, speakers' socioeconomic status possibly should take into account, in addition to their income and level of education, their type of residence, occupation, and access to cultural goods. The corpus will also be stratified according to speakers' generation in the city, in order to examine the contribution of different groups of migrants and immigrants in the community, and speakers' area of residence, which is also an index of socioeconomic status.

During this first phase of the project, information on these variables is collected through the sociolinguistic interview and post-recording questionnaires, which will enable preliminary analyses of their role in the sociolinguistic stratification in São Paulo.

Speakers to be recorded have been contacted through the "friend of a friend" method (Milroy, 2004). Our experience has shown that speakers in the city are very resistant to talking to a "stranger" (the researcher); however, when introduced by a common acquaintance, speakers tend to be much more receptive and solicitous, a fact that also has consequences for naturalness of speech. After a speaker has been recorded, the researcher asks her/him to suggest another speaker. As a means to ensure that informants do not belong to the same or few social networks, the new suggested speaker can only be recorded if he or she is not acquainted with the person who indicated the current informant. For instance, in the example in Figure 2, B has indicated two new speakers, C and D, but only the latter can be selected as a new informant.

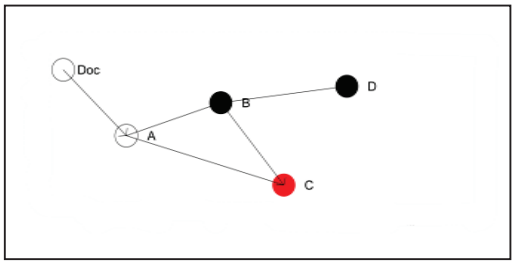

Figure 2: Selection of informants

The interview schedule has the twofold objective of obtaining samples of spontaneous speech by Paulistanos of varied sociolinguistic profiles and more information on these speakers' living conditions, sociolinguistic evaluations and perceptions (Labov, 2006). It is divided into two parts. The first one is more personal and covers topics such as the speakers' neighborhood, childhood, parents and family, education, current occupation, social network, and leisure activities. It aims at obtaining narratives in the past (e.g. "What was your childhood like in neighborhood X?"), in the present (e.g. "In your leisure time, what do you and your family like to do?") and in the future (e.g. "What would you do if you won the lottery?"), as well as opinion accounts (e.g. "What do you think of the new law for gay marriage?"). The second part contains more specific questions about the speakers' relation to the city and their perceptions on Paulistano identities (e.g. "When you were in (another city), did people recognize you as a Paulistano? If so, how?"). In the last part of the interview, speakers are asked to read a word list, a news report, and a 'statement' (a text with strong marks of oral language). Although the interview schedule is divided into two parts, it enables easy transition between topics and has yielded natural sounding conversations.

After the interview is recorded, the fieldworker fills out a form with detailed speaker's sociolinguistic information (date of birth, occupation, family's place origin and first generation that migrated to São Paulo, schools, place(s) of residence etc.), and makes note of any relevant additional information in the fieldwork journal. The informant is also asked to fill out a socioeconomic form, if he/she feels comfortable to do so, containing seven multiple-choice questions about their monthly income and living conditions. Our experience has shown that the multiple-choice form greatly improves the chance of obtaining these data (instead of having the informant orally answer these questions directly to the fieldworker).

Each sociolinguistic interview is about 60-70 minute long and has been stored in .wav (stereo, 44,100 Hz) format. The recordings have been made with TASCAM DR-100 recorders and two Sennheiser HMD26 microphones (one for the fieldworker and one for the informant). Although it could be argued that the presence of these technical paraphernalia possibly enhances the Observer's Paradox (Labov, 2006), we find that speakers' occasional uneasiness tends to decrease considerably after some 15 minutes of recording and, more importantly, that the improved audio quality is worth the trouble, especially in a city as noisy as São Paulo.

All interviews are then evaluated by four members of the research group not involved in the field recordings, according to the speakers' fitness to the sociolinguistic profile, audio quality, naturalness of conversation, and conformity to the interview schedule. The 82 previously collected interviews during the pilot experience have also been evaluated according to the same parameters, and some of them may be included in the final corpus to be made available online, in addition to the 60 recordings of the present data collection phase, as long as they meet the high-quality requirements.

The criteria for transcribing the recordings follow a simplified semiortographic approach in order to make the material more easily available in a written medium. The following criteria aim at facilitating the manipulation of text files in softwares such as R (Gries, 2009; Hornik, 2011) to automatically identify and extract tokens of a variable into a spreadsheet program (Oushiro, 2012).

Transcripts do not contain any special formatting such as boldface, italics, tab stops, columns, and are saved in plain text (.txt) with UTF-8 encoding. Orthographic rules of Brazilian Portuguese are followed in every case, even if speakers produce variants that differ from the written standard. The idea here is that a transcriber is unable to pay attention to all variable phenomena simultaneously – e.g. monophthongization of /ow, ej/, diphthongization of nasal /e/, postvocalic /r/ deletion, nasal assimilation of /ndo/, vowel raising of unstressed /e,o/, to name a few. In addition to creating unintelligible texts, this would probably cause transcripts to be unstandardized; further, the fact that the recordings will be made available lessens the need for a highly detailed transcript. On the other hand, grammatical variables should not be "corrected" by the transcriber (e.g. lack of

nominal agreement). Punctuation is limited to ellipses (to signal pauses), and question and exclamation marks (to indicate intonation of certain phrases). Capital letters are only employed in proper names (e.g. cities, institutions), abbreviations (e.g. USP, and identifying speakers (e.g. S1, D1).

GESOL-USP has also been developing parallel data collection projects, in addition to gathering a sample from the community at large. These parallel projects and studies are centered on specific groups of speakers and/or social variables within the city: residents of the upper class neighborhood Itaim Bibi (Ciancio, 2012); social class (Faria, 2012); gay men and gender (Soriano, 2012); different groups of migrants – Paraibanos (Mendes, forth) and Alagoanos (Silva, 2012). These studies aim at describing and contrasting general sociolinguistic patterns of the community and their uses within certain social groups residing in the city.

Based on the corpus collected so far, the research group has been developing studies of sociolinguistic variation in Paulistano Portuguese: the variable realization of coda (-r) as a tap or a retroflex, in words such as *porta* 'door' and *mulher* 'woman' (Mendes, 2009, 2010; Mendes & Oushiro forth); variable nasal (e) as a monophthong or a diphthong, in words such as *fazenda* 'farm' (Mendes, 2010; Oushiro, 2011); verbal negative structures (e.g. *Não vou* vs. *Não vou não* 'I won't go') (Rocha, 2012); nominal and verbal agreement (Silva, 2012; Oushiro, 2011).

3. Conclusion

The SP-2010 Project has been collecting a contemporary corpus of Paulistano Portuguese and fostering the development of sociolinguistic studies focusing on the correlations between variable linguistic uses and social identities. By 2013, more than 60 sociolinguistic interviews (audio and transcriptions) will be made available online to the linguistic community. Parallel to this data collection project, a number of studies have also been analyzing specific social networks and communities of practice in the city, in contrast with larger community variational patterns, as to provide a broader and more detailed description of linguistic uses in São Paulo.

4. Acknowledgements

This research is funded by FAPESP (Grant 2011/09278-6).

5. References

Bisol, L., Menon, O.P.S. and Tasca, M. (s/d). VARSUL, um banco de dados. Available at: <http://www.varsul.org.br/?modulo=secao&id=9>.

Bourdieu, P. (1991) *Language and symbolic power*. Cambridge: Polity Press.

Castilho, A.T., Preti, D. (Eds.). (1986). *A linguagem falada culta na cidade de São Paulo: materiais para seu estudo,* vol. I – Elocuções Formais. São Paulo: T.A. Queiroz.

Castilho, A.T., Preti, D. (Eds.). (1987). *A linguagem falada culta na cidade de São Paulo: materiais para seu estudo*, vol. II – Diálogos entre dois informantes. São Paulo: T.A. Queiroz/FAPESP.

Castilho, A.T. (2007). Projeto para a História do Português Paulista. Projeto Temático de Equipe (Proc. FAPESP n. 06/55944-0).

Chambers, J.K. (1995). *Sociolinguistic theory: linguistic variation and its social significance*. Oxford: Blackwell.

Cheshire, J. (2004). Sex and gender in variationist research. In J.K. Chambers, P. Trudgill & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change*. Oxford: Blackwell.

Ciancio, R. (2012). Estudo sociolinguístico da fala paulistana por falantes do Itaim Bibi. Research Project.

Coelho, R.F. (2006). É nóis na fita! Duas variáveis linguísticas numa vizinhança da periferia paulistana. O pronome de primeira pessoa do plural e a marcação de plural no verbo. Master's dissertation.

Eckert, P. (2005). Three waves of variation study: the emergence of meaning in the study of variation. Manuscript, s/d. Available at: <www.stanford.edu/~eckert/PDF/ThreeWavesofVariation.pdf>.

Faria, C.B. de (2012). Para a inclusão de "classe social" nos estudos sociolinguísticos em São Paulo. Research Project.

Gonçalves, S.C.L. (2003). O português falado na região de São José do Rio Preto: constituição de um banco de dados anotado para o seu estudo. Research project. Available at: <www.iboruna.ibilce.unesp.br/historico/Projeto>.

Gries, S.Th. (2009). *Quantitative Corpus Linguistics with R*. NewYork: Routledge.

Hora, D. da (Ed.). (2004). *Estudos Sociolinguísticos*: perfil de uma comunidade. Santa Maria: Palotti.

Hornik, K. (2011). R FAQ. Available at : <http://cran.r-project.org/doc/FAQ/R-FAQ.html>.

IPEA (2011). Comunicados do IPEA no. 115 – Perfil dos migrantes em São Paulo, 2011. Available at: <www.ipea.gov.br/portal/images/stories/PDFs/comunicado/111006_comunicadoipea115.pdf>.

Labov, W. (1984). Field methods of the Project on Linguistic Change and Variation. In J. Baugh, J. Sherzer (Eds.), *Language in Use: Readings in Sociolinguistics*. Englewood Cliffs: Prentice Hall, pp. 28--54.

Labov, W. (2001). *Principles of linguistic change*: social factors. Oxford & Cambrige: Blackwell.

Labov, W. (2006). *The Social Stratification of English in New York City*. Cambridge: Cambridge University Press.

Mendes, R.B. (2009). Who sounds /r/-ful? The pronunciation of coda /-r/ in the city of São Paulo. *Paper presented at NWAV38*. University of Ottawa.

Mendes, R.B. (2010). Sounding Paulistano: variation and correlation in São Paulo. *Paper presented at New Ways of Analyzing Variation - NWAV39*. University of Texas at San Antonio.

Mendes, R.B. (2011). SP-2010 – Construção de uma amostra da fala paulistana. Projeto de Pesquisa.

Mendes, R.B. (forth). A pronúncia retroflexa do /-r/ na fala paulistana. In D. Hora, E.V. Negrão (Eds.), *Estudos da Linguagem. Casamento entre temas e perspectiva*. João Pessoa: Ideia/Editora Universitária, pp. 282--299.

Mendes, R.B., Oushiro, L. (forth). Percepções sociolinguísticas sobre as variantes tepe e retroflexa na cidade de São Paulo. In D. Hora, E.V. Negrão (Eds.), *Estudos da Linguagem. Casamento entre temas e perspectiva*. João Pessoa: Ideia/Editora Universitária, pp. 262--281.

Milroy, L. (2004). Social networks. In J.K. Chambers, P. Trudgill and N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change*. Oxford: Blackwell.

Oushiro, L. (2011). Identidade na pluralidade: produção e percepção linguística na cidade de São Paulo. Research project.

Oushiro, L. (2012). Analyzing (-r) with R. Paper presented at the 2012 GSCP International Conference. Available at: <www.letras.ufmg.br/gscp2012-eng/data1/arquivos/84.pdf>.

Paiva, M.C., Scherre, M.M.P. (1999). Retrospectiva sociolinguística: contribuições do PEUL. In *DELTA* (online), vol. 15, n.spe., pp. 201--232.

Poplack, S. (1989). The care and handling of a mega-corpus: the Ottawa-Hull French Project. In R. Fasold, D. Schiffrin (Eds.), *Language Change and Variation*. Amsterdam: Benjamins, pp. 411--451.

Preti, D., Urbano, H. (Eds.). (1988). *A linguagem falada culta na cidade de São Paulo*: *materiais para seu estudo*, vol. III – Entrevistas. São Paulo: T.A. Queiroz/FAPESP.

Preti, D., Urbano, H. (Eds.). (1990). *A linguagem falada culta na cidade de São Paulo: materiais para seu estudo*, vol. IV – Estudos. São Paulo: T.A. Queiroz/FAPESP.

Rocha, R.S. (2011). Negação pós-verbal no português paulistano: definição do envelope de variação. In *Seminário do GEL, 59, Programação*, Bauru: GEL, 2011. Available at: <http://gel.org.br/detalheResumo.php?trabalho=7564>.

Rodrigues, A.C.S. (1987). A concordância verbal no português popular em São Paulo. PhD Thesis. FFLCH-USP.

Rodrigues, A.C.S. (2009). Fotografia sociolinguística do português do Brasil: o português popular em São Paulo. In A.T. Castilho (Ed.), *História do Português Paulista*. Campinas: Instituto de Estudos da Linguagem/UNICAMP.

Silva, F.G. (2012). Concordância nominal: um contraste dentro da cidade de São Paulo. Research project.

Soriano, L. (2012). Estudo sociolinguístico de gays paulistanos em diferentes situações de fala. Research project.

Tagliamonte, S. (2006) *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.