Letrônica, Porto Alegre, v. 6, n. 1, p. 319-337, jan./jun., 2013

O QUE É QUE SE FAZ COM OS RESULTADOS DO VARBRUL?

WHAT CAN BE DONE WITH THE RESULTS OF VARBRUL?

Odete Pereira da Silva Menon[*]
Loremi Loregian-Penkal[**]
Edson Domingos Fagundes[***]

Resumo: Este trabalho objetiva trazer à tona algumas discussões sobre a leitura dos dados estatísticos e as conclusões tiradas *a priori*, evidenciando a importância de se considerar, entre outras questões, a ortogonalidade na distribuição dos dados, pois alguns dos resultados que obtivemos com o programa VARBRUL geraram algumas reflexões. Para isso, vamos apresentar (*i*) resultados obtidos na análise da alternância *tu/você* nas localidades do VARSUL em que ela se apresenta, e (*ii*) da análise da Concordância Nominal (CN) nas cidades paranaenses do banco – Londrina, Irati e Pato Branco. Na análise da CN, constatamos que há diferenças no tocante à seleção de variáveis, a depender do tipo de colonização da região analisada (para Londrina e Pato Branco, colonização interna; Irati, colonização externa), razão pela qual deixamos de considerar, neste trabalho, a cidade de Curitiba, que representa o Paraná "velho". Na análise da alternância *tu/você* (LOREGIAN-PENKAL, 2004) ficou evidenciada a importância de se considerar a análise da variação no indivíduo o que, *mutatis mutandis*, pode se aplicar aos resultados individualizados da CN nas três cidades. As análises dos dados foram feitas pelo viés da sociolinguística quantitativa, pelo pacote VARBRUL, e levaram em conta fatores linguísticos e extralinguísticos. Assim esperamos demonstrar como o banco de dados VARSUL vem contribuindo para a descrição dos fenômenos variáveis do português do Brasil.
Palavras-chave: Sociolinguística Quantitativa; Ortogonalidade; Alternância Tu/Você; Concordância nominal; Ocupação do território; Banco de dados VARSUL.

Abstract: This work aims at bringing up some discussion about the analysis of statistical data, and a priori conclusions, highlighting the importance of considering, among other issues, the orthogonality in the distribution of data, since some reflections were generated by some of the results obtained with the VARBRUL program. Therefore, we will present the (i) results from the analysis of tu/você switch in the localities of VARSUL where the process exists, and (ii) analysis of Noun Agreement (NA) in the Paraná cities - Londrina, Irati and Pato Branco. In the analysis of NA, we realized that there are differences regarding the selection of variables, depending on the type of colonization of the analyzed region (for Londrina and Pato Branco, internal colonization; for Irati, external colonization) - the reason why Curitiba city, the "old" Paraná, was not considered in this paper. In the analysis of tu/você switch (LOREGIAN-PENKAL, 2004), we emphasize the importance of considering the analysis of the variation in the individual which, mutatis mutandis, can be applied to

[*] UFPR – Universidade Federal do Paraná/CNPq. Programa de Pós-Graduação em Letras. Curitiba. Paraná. Brasil. 80410-000. odete@ufpr.br.
[**] UNICENTRO – Universidade Estadual do Centro-Oeste do Paraná/CNPq. Departamento de Letras. Campus de Irati. Paraná. Brasil. 84500-000. llpenkal@irati.unicentro.br.
[***] UTFPR – Universidade Tecnológica Federal do Paraná. Departamento de Línguas Estrangeiras Modernas. Campus Curitiba. Paraná. Brasil. 80230-190. edsondfagundes@utfpr.edu.br.

individual results of the NA in the three cities. The data analyzes followed the quantitative sociolinguistics steps with VARBRUL package, and took into account extralinguistic and linguistic factors. We hope to demonstrate how the VARSUL database has been contributing to the description of the variable phenomena in Brazilian Portuguese.
Keywords: Quantitative Sociolinguistics; Orthogonality; Tu/você switching; Noun agreement; Occupation of territory; VARSUL database.

Considerações iniciais

Neste trabalho, discutimos a leitura que se tem feito dos dados estatísticos obtidos a partir das rodadas estatísticas com o programa VARBRUL. O foco da discussão está centrado na análise de resultado de pesquisas feitas no âmbito do Projeto VARSUL e em um possível conservadorismo observado em Irati em relação às outras cidades do banco no Paraná. Nesse aspecto, há algumas questões que precisamos considerar. Dentre elas está (*i*) uma pequena quantidade de dados não auxilia a análise que o programa estatístico pode nos fornecer; porque a ortogonalidade é justamente o equilíbrio na distribuição dos dados nos diferentes grupos de fatores que compõem uma determinada amostra; ora, uma amostra muito pequena pode não preencher todas as células de maneira equilibrada; (*ii*) nem sempre a amostra é homogênea em sua constituição, pois pode acontecer de termos informantes que para determinado fenômeno sejam categóricos no tocante à realização de uma das variantes da variável em estudo. Nesse caso, o indicado seria fazer um refinamento da análise, considerando o indivíduo, obtendo, assim, uma descrição mais confiável. De modo semelhante, ao analisarmos os dados de diferentes cidades e compará-los, adotamos os mesmos procedimentos, ou seja, analisar o conjunto dos dados e levantar hipóteses a partir daí ou refinar a análise considerando também o comportamento de cada cidade.

A fim de demonstrar algumas dessas situações aqui levantadas, vamos nos pautar na retomada da análise de alguns dos resultados obtidos em pesquisas feitas pelo grupo VARSUL-Paraná no tocante à análise da concordância nominal (CN) e da alternância TU/VOCÊ.

Antes de tratarmos do tema que motiva este trabalho, apresentaremos um breve histórico a respeito do Projeto VARSUL.

1 Breve histórico do Projeto VARSUL

O Banco de dados VARSUL[1] é o resultante da execução do projeto *Variação Linguística Urbana na Região Sul do Brasil*, cuja concepção foi idealizada em 1984, por Leda Bisol, que reuniu alguns pesquisadores em Porto Alegre. O projeto proposto pela pesquisadora deveria se espelhar no projeto pioneiro de levantamento sociolinguístico no Brasil: Projeto Censo Linguístico do Rio de Janeiro, coordenado por Anthony Julius Naro, e executado no final dos anos 70, na Universidade Federal do Rio de Janeiro (UFRJ), com os primeiros resultados publicados no início dos anos 80. O Projeto Censo limitou a coleta de dados à cidade do Rio de Janeiro, realizando as entrevistas em diferentes bairros, representativos das diferentes comunidades cariocas, sobretudo do ponto de vista social.

Embora o modelo de coleta de dados fosse o do Censo, para dar conta da diversidade étnica da região, chegou-se a um consenso: não bastaria pesquisar as capitais dos três estados (Paraná, Santa Catarina e Rio Grande do Sul); seria necessário incluir algumas das etnias representativas da ocupação étnica diferenciada não só da região Sul, mas de cada estado individualmente. A razão disso estava no fato de que se pretendia pesquisar se o português da região sul diferiria dos demais dialetos do PB (português brasileiro) como consequência do povoamento distinto desses estados (essa região era praticamente despovoada no tempo em que mais entraram escravos no Brasil). Por isso, em algumas das localidades, foram entrevistados informantes bilíngues; nas demais, informantes monolíngues de português.

Outra diferença em relação ao Projeto Censo diz respeito ao sistema de transcrição de dados e ao armazenamento das entrevistas transcritas. No Rio, o armazenamento foi feito no grande computador da universidade, o que posteriormente se revelou extremamente dificultoso no que diz respeito ao acesso aos dados pelos pesquisadores. No VARSUL, optou-se por um sistema que possibilitasse o armazenamento em microcomputadores, facilitando o acesso aos dados. A transcrição dos dados da região Sul, pelas suas próprias características diferenciadas, exigia um sistema de indicação de idiossincrasias, sobretudo relativas à pronúncia: daí ter-se

[1] Além dos trabalhos já feitos que apresentam a descrição e memória do Projeto VARSUL, o grupo de pesquisa do VARSUL Paraná apresentou trabalho no Congresso Internacional *Proceedings in Language and Text Corpus Design and Linguistic Corpus Analysis*, cuja publicação também está disponível em http://www.linguistik-online.de/38_09/menonEtAl.html.

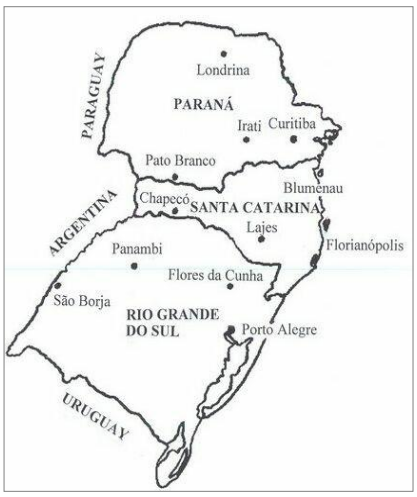
escolhido um sistema de transcrição em três linhas: a primeira linha registra a sintaxe real da fala dos informantes, as hesitações e interrupções; a segunda linha, as pausas e aspectos fonéticos variáveis e a terceira linha, a classificação morfossintática e a marcação de aspectos prosódicos, como ênfase e velocidade de fala.

Com respeito à metodologia do levantamento de dados, seguiu-se a linha laboviana, inspirando-se a transcrição das entrevistas no trabalho realizado pela equipe do Projeto Censo do Rio de Janeiro, porém já configurada para ser armazenada em microcomputadores.

O projeto VARSUL foi concebido com o objetivo de instalar, a curto prazo, um banco de dados linguísticos que possibilitasse, mais adiante, a descrição da variedade linguística urbana da Região Sul do Brasil e de suas sub-variedades locais.

A implantação desse banco foi feita inicialmente com os dados das capitais e de uma cidade do interior de cada estado. Numa segunda etapa, esses dados foram complementados para cobrir as áreas urbanas mais representativas das regiões que sob os pontos de vista histórico, social e cultural eram salientes e relevantes em cada um dos estados da Região Sul do Brasil. Assim, em cada estado foram selecionados quatro municípios representativos de grupos populacionais comprovadamente relevantes no seu processo de ocupação. Assim, como pode ser conferido na figura 1, para o estado do Rio Grande do Sul as cidades contempladas foram: Porto Alegre; Flores da Cunha; Panambi e São Borja. Em Santa Catarina as cidades foram: Florianópolis; Blumenau; Lages e Chapecó. No Paraná as quatro cidades selecionadas foram: Curitiba; Irati; Londrina e Pato Branco.

Figura 1 – Distribuição geográfica das localidades incluídas na amostra



1.1 O VARSUL no Paraná

O estado do Paraná apresenta um panorama linguístico extremamente variado: a razão principal dessa multiplicidade está nas diversas origens da população paranaense, que se constituiu de diferentes levas de grupos populacionais: o colonizador português dos primeiros séculos, os imigrantes europeus e asiáticos (séculos XIX e XX) e os migrantes brasileiros das últimas décadas, vindos principalmente de Minas, São Paulo e Rio Grande do Sul. São, portanto, várias as modalidades do português faladas no estado.

Como não há levantamento sistemático dessas variedades, no projeto se tentou dar conta dessa ocupação étnica considerando as seguintes áreas: o Norte do estado foi povoado por mineiros e paulistas durante a expansão da agricultura cafeeira (anos 1930), sendo a cidade de Londrina a que representa essa ocupação do território. No Sudoeste e Oeste, o linguajar trazido pelos colonos gaúchos e catarinenses descendentes de gaúchos, responsáveis pela ocupação agrícola daquela parte do estado, vai ser representado pela cidade de Pato Branco.

Na região em que se concentrou a imigração de povos eslavos (russos, poloneses e ucranianos) e que ainda se conserva parcialmente bilíngue, a cidade de Prudentópolis (ucranianos) seria mais representativa, mas, por não ser núcleo urbano suficientemente estratificado para a amostra, foi preterida em favor de Irati (poloneses e ucranianos), cidade maior (embora ali a maioria da população não seja bilíngue). No Centro-Sul, também chamado Paraná Velho, que mais individualiza o estado do ponto de vista linguístico, foi selecionada Curitiba, capital do estado.

Para a constituição da amostra, levaram-se em consideração as seguintes características sociais comprovadamente significativas em pesquisas sociolinguísticas anteriores: sexo (masculino e feminino); idade (25-45 e acima de 50 anos) e escolaridade (primário, ginasial e segundo grau).

Definiu-se que cada município deveria ser representado na amostra por um conjunto de 24 entrevistas, correspondentes a 12 perfis (2 sexos x 3 níveis de escolaridade x 2 faixas etárias), cada um representado por dois entrevistados. Com a definição desses perfis, buscou-se localizar informantes em diferentes bairros com população permanente considerável.

Além disso, os falantes tinham que, necessariamente, preencher os seguintes pré-

requisitos: (*i*) falar apenas português (exigência para os entrevistados nas capitais, mas não nas áreas bilíngues); (*ii*) ter morado na cidade pelo menos 2/3 de sua vida; (*iii*) não ter morado fora da região por mais de um ano no período de aquisição da língua nativa (2 a 12 anos); (*iv*) não causar estranheza a outros moradores da região.

Não foram incluídos na etapa inicial os analfabetos e universitários pelo fato de constituírem população alvo de estudos dialetais (analfabetos) e da norma urbana regional culta (universitários).

A faixa etária abaixo de 20 anos não foi considerada por não apresentar a consistência linguística necessária para os objetivos do estudo inicial (idiossincrasias frequentes). A decisão de não incluir essa faixa etária levou em conta também o tamanho da amostra. Como o banco devia incluir 4 municípios em cada estado, o número de informantes de cada município foi reduzido, considerando-se apenas as características sociais que se mostraram comprovadamente significativas em estudos anteriores.

Vale salientar que o Banco VARSUL vem sendo ampliado[2], com o acréscimo de novas amostras. À amostra básica, constituída de informantes distribuídos por três graus de escolaridade, sexo, faixa etária, outras vêm sendo acrescidas: mais uma faixa etária (15-24 anos) e mais um nível de escolaridade (universitários), esta última só nas capitais. Note-se também que o VARSUL vem se tornando um lugar privilegiado na formação de novos pesquisadores, abrindo portas a alunos de graduação (bolsistas de iniciação científica), mestrandos, doutorandos e pós-doutorandos.

2 Preliminares ao estudo da CN no Paraná

No estudo sobre as ocorrências do modo subjuntivo (MS), com objetivo de testar se havia alternância entre esse modo e o indicativo (MI) nos contextos de oração independente e de oração subordinada, Fagundes (2007) considerou *modo verbal* a variável dependente e *tipo de oração, tempo verbal da oração principal*, *tempo verbal da ocorrência* (na oração *subordinada* ou na *independente*) e *modalidade* variáveis linguísticas. As variáveis sociais foram *sexo*, *faixa etária*, *grau de escolaridade* e *etnia*, visto utilizar as 4 cidades do banco VARSUL do Paraná. O número de dados nos

[2] Para mais informações a respeito do projeto, das pesquisas realizadas e em andamento, consulte a página do Projeto VARSUL: www.varsul.org.br.

contextos em que se esperava ocorrer a alternância entre os dois modos verbais totalizou 2.718, sendo 2.434 de subjuntivo e 284 de indicativo.

As rodadas revelaram que os fatores intervenientes, que condicionam as escolhas e a alternância de uso dos modos verbais feitas pelo falante, são de ordem social – no caso das *cidades* – e não social – no caso do *tipo de oração* e da *modalidade*. As variáveis não sociais pertencem ao sistema e a alternância detectada, portanto, somente ocorre onde ele permite.

Os resultados apontaram para Londrina uma indefinição quanto à escolha e ao emprego dos modos feito por parte dos falantes (0,51 *modo subjuntivo*, 0,49 *modo indicativo*); Curitiba (0,44 *modo subjuntivo*, 0,56 *modo indicativo*) e Pato Branco (0,46 *modo subjuntivo*, 0,54 *modo indicativo*) apresentaram uma alternância no uso dos modos verbais, indicando que se encontram em um estágio mais avançado no que se refere à alternância de uso do MI e MS e que a distinção entre as duas cidades pode ser observada com mais nitidez no grupo de fatores tipo de oração; Irati, além de se distinguir de Curitiba na distribuição dos dados no grupo de fatores modalidade, é a cidade que apresenta um perfil mais conservador se considerarmos que a alternância entre os modos verbais pode ser um fenômeno inovador (0,62 *modo subjuntivo*, 0,38 *modo indicativo*).

Bandeira (2007) analisou o pronome *se* – indeterminador, reflexivo, recíproco, enfático – nas quatro cidades do VARSUL Paraná (Curitiba, Irati, Londrina e Pato Branco) a fim de estudar o apagamento desse pronome na função de sujeito. Para isso testou os 3.829 dados levantados em 12 variáveis independentes (8 linguísticas e 4 sociais) na perspectiva da sociolinguística quantitativa (LABOV, 1972). O resultado percentual de apagamento foi 45 por cento. A partir das rodadas com os grupos de fatores, os pesos relativos (P.R.) revelaram que a ausência do pronome se manifestou mais no tipo *enfático* (0,94 em Irati e 0,95 em Pato Branco), seguido do *indeterminador* (0,79 em Pato Branco e 0,70 em Curitiba). Em contrapartida, os mais resistentes ao apagamento são o *reflexivo* e o *recíproco*. No tocante às localidades, a cidade de Irati é a mais conservadora quando se trata da manutenção do pronome: o peso relativo de ausência geral desse pronome na cidade é só de 0,25; seguido de Pato Branco com 0,45, Curitiba com 0,67 e Londrina com 0,71. Como podemos observar, esses resultados parecem apontar um comportamento diferenciado das localidades no tocante à

ocupação étnica e à diversidade populacional. Os demais resultados de Bandeira (2007), relativos às variáveis sociais, mostraram que a *faixa etária* só foi selecionada para Irati e Pato Branco, cidades em que os mais jovens apagam mais o pronome (0,73 em Irati e 0,56 em Pato Branco); *sexo* só foi selecionada em Pato Branco com 0,59 de ausência para o feminino. *Escolaridade* parece não ser relevante para o fenômeno estudado por Bandeira, visto não ter sido selecionada em nenhuma das cidades.

A partir desses dois trabalhos, a equipe do grupo de pesquisa VARSUL/Paraná resolveu estudar outro fenômeno morfossintático para testar se realmente Irati é a mais conservadora dentre as cidades do banco VARSUL do Paraná. Já que Irati tinha se revelado diferente de Curitiba em alguns resultados inesperados, resolvemos verificar um fenômeno característico da região Sul e que, aparentemente, contrariava a regra proposta por Scherre (1988) e outros trabalhos daí decorrentes: a concordância nominal com o pronome possessivo. Além disso, havia somente dois trabalhos enfocando a CN na região Sul: Dias (1996) e Fernandes (1996).

Dias trata da CN nos predicativos e particípios passivos em três cidades — Florianópolis (açorianos), Chapecó (italianos) e Irati (eslavos) —, com objetivo de testar a variável etnia e chega à conclusão de que o grupo dos italianos (0,57) marca mais o plural, seguidos dos eslavos (0,48) e por último os açorianos (0,44). (DIAS, 1996, p. 91 e p. 106).

Fernandes estudou a CN nos sintagmas nominais nos informantes do VARSUL de Florianópolis, Chapecó, Panambi e Irati, para as situações informais. Os resultados de Fernandes (1996, p. 99) apontam para uma maior concordância nos grupos alemão (0,55) e eslavo (0,55) e menor índice de concordância entre açorianos (0,47) e italianos (0,44).

Além disso, a pesquisadora gravou 19 informantes em situações que ela considerou formais: programas esportivos em televisão, entrevistas televisivas e defesas de dissertação de Mestrado, criando, assim, quatro categorias de formalidade: informal 1 (as entrevistas VARSUL); informal 2 (comentários esportivos em TV); formal 1 (entrevistas televisivas); formal 2 (defesa de dissertação de mestrado). Os resultados nesse grupo de fatores apresentaram um *continuum* de menor concordância na situação informal 1 (0,43) para maior concordância na formal 2 (0,82), que indicam claramente a tendência de que quanto mais formalidade da situação, mais o falante "capricha" na

aplicação da regra; os níveis intermediários informal 2 (0,56) e formal 1 (0,65) favorecem levemente a regra, num crescendo.

Como vemos, apesar de os grupos de fatores e a seleção das cidades não serem os mesmos em Dias, e em Fernandes, no tocante à variável etnia, os grupos açoriano e eslavo apresentam semelhanças na aplicação de ambas as regras de concordância e o grupo dos italianos favorece a regra nos predicativos (0,57), mas desfavorece a regra nos sintagmas (0,44), embora o distanciamento seja quase o mesmo, para cima ou para baixo do ponto neutro. Segundo a teoria, isso demonstra que a variável etnia ao ser favorecida em 7 pontos ou desfavorecida em 6 pontos de peso relativo, portanto, com uma diferença de 0,13 entre uma variante e outra (que deve ser considerada na análise dos dados); no que se refere à relação com o ponto neutro (0,50), ambas são pouco significativas.

Assim, mesmo dispondo desses resultados para Irati, resolvemos refinar os grupos de fatores visto que em Fernandes (1996) não foi enfocado o caso específico da concordância do possessivo.

2.1 A pesquisa em Irati e a extensão às demais cidades do Paraná

Inicialmente, o levantamento de dados em Irati tinha como objetivo testar o comportamento dos falantes dessa cidade na aplicação da regra de concordância nominal intrassintagmática a fim de verificar se, nesse fenômeno, persistia o caráter "mais conservador" revelado pelos trabalhos de Fagundes (2007) e Bandeira (2007). O motivador da análise orientava-se na aplicação da regra de concordância proposta por Scherre (1988), a de que "quanto mais à esquerda do núcleo, maiores as probabilidades de aparecer marca de concordância", com relação à ocorrência do pronome possessivo na região sul. Nesta região, não soa estranho dizer (ou ouvir) **o meus filho, a minhas tias, a nossas viatura,** ao lado de ocorrências canônicas como **as tuas orações**. Para Scherre, essa construção trazia problemas e várias vezes (em assessorias ao projeto VARSUL ou em cursos ministrados em eventos) ela indagou da questão de gramaticalidade de tal construção. No entanto, falantes tanto do Paraná, quanto de Santa Catarina ou do Rio Grande do Sul, reiteravam a "normalidade" dessas ocorrências.

Feita a seleção dos dados de CN intrassintagmática em Irati, de acordo com os grupos de fatores comuns aos utilizados em outros trabalhos sobre o assunto (presença/ausência de marca; posição linear do elemento no sintagma nominal; classe gramatical da palavra; relação com o núcleo; marcas precedentes; processos morfológicos de formação de plural; tonicidade do item no singular; traço semântico/animacidade do item; grau do substantivo ou adjetivo; contexto fonológico seguinte; sexo; idade; escolaridade; cidades)[3], fizemos rodadas iniciais.

Dos treze grupos de fatores independentes, foram selecionados nove, pela ordem: relação com o núcleo, posição linear, escolaridade, grau, formação do plural, sexo, animacidade, idade e classe gramatical. Com relação ao comportamento do possessivo, os SNs. podiam apresentar concordância total, como em (01) a (03):

(1) os meus filhos (IRT 05 726)[4]
(2) as minhas tias (PBR 03 399)
(3) os meus netos (LDN 10 694)

e aparente discordância com a regra proposta por Scherre, como nos exemplos (4) a (8).

(4) **do** meus **ermão** (IRT 16 843)
(5) **do** meus **filho** (PBR 02 1000)
(6) **o** meus **parente** (PBR 05 062)
(7) **a** minhas **criança** (LDN 07 1143)
(8) **a** nossas **viatura** (LDN 15 028)

Chamamos de aparente discordância porque, na análise dos SNs, o artigo aparece "sem" a marca de plural, como em ***o*** *meus filho*. Analisando a recorrência dessa estrutura nos dados e fatos da prosódia da língua portuguesa, construímos a hipótese de que o

[3] Procuramos utilizar os mesmos Grupos de Fatores (doravante GFs) já empregados em outras pesquisas, a fim de se cumprir o requisito sempre apontado por Gregory Guy (em intervenções em congressos) da comparabilidade dos resultados, ou seja, para que se possa comparar, realmente, resultados, é necessário que os GFs. sejam exatamente os mesmos. Esse critério se revelou de difícil execução, porque existem variações regionais diferenciadas, como a ocorrência de mininim, p. ex., em amostras mineira e brasiliense, e o caso dos possessivos na região sul.

[4] Leia-se: cidades (IRT Irati, LDN Londrina, PBR Pato Branco); número do informante e número da linha em que o dado foi retirado.

artigo constitui a primeira sílaba daquilo que chamamos *vocábulo fonológico*[5]: o que antecede o núcleo *filho* é o determinante ***o-meus***, que carrega a marca de plural. Portanto, aquilo que parece fugir da regra, na realidade confirma a regra, visto que é o elemento que está mais à esquerda do núcleo o portador da marca de plural.

De posse dos resultados de Irati, passamos a levantar os dados da segunda cidade do córpus VARSUL Paraná, Pato Branco, colonizada por gaúchos à procura de terras mais baratas e abundantes (o que já não havia mais no RS, por causa da explosão demográfica: sucessivas divisões de glebas por descendentes numerosos), assim como por "catarinenses", isto é, descendentes de gaúchos já estabelecidos no oeste de SC em levas anteriores de migração. Fizemos a codificação dos dados com exatamente os mesmo GFs. usados em Irati e os dois primeiros grupos selecionados pelo programa estatístico foram os mesmos, mas inversamente posicionados, o que nos levou a indagar se eles não estariam se sobrepondo. Além disso, o número de fatores selecionados como relevantes na aplicação da regra em Pato Branco foi menor que o de Irati, pois ficaram de fora processos de formação do plural e classe gramatical, discussão que levamos a cabo em texto publicado na revista Gragoatá (MENON, FAGUNDES; LOREGIAN-PENKAL, 2010). Os demais GFs., coincidentes em ambas as localidades, não tiveram a mesma ordem de seleção. *Escolaridade*, dentre eles, foi o que mais chamou a atenção, por vir em último lugar, quando, em Irati, havia sido selecionado em terceiro lugar.

Começavam, então, a se delinear diferenças entre as cidades, o que, em parte, viria a confirmar a hipótese de que *etnia* seria fator relevante nas diferenças dialetais. Passamos, então, aos dados de Londrina.

Aplicadas as mesmas etapas de levantamento de dados, codificação e rodadas estatísticas, obtivemos o seguinte resultado para Londrina: se em Irati e Pato Branco as duas primeiras variáveis foram idênticas, privilegiando a *posição* e a *relação com o núcleo*, tal qual a premissa de Scherre, Londrina apresenta em primeiro lugar a variável *marcas precedentes*, seguida de *classe gramatical*. Ora, em Pato Branco, essas duas variáveis sequer são selecionadas como relevantes para a CN; em Irati, *classe gramatical* foi a última selecionada e *marcas precedentes* não foi considerada relevante pelo programa. Diferentemente de Londrina, Irati e Pato Branco selecionaram *grau do*

[5] Estamos preparando texto em que analisamos as diferentes possibilidades de CN com o possessivo, na perspectiva de analisar o fenômeno como vocábulo fonológico (MENON, LOREGIAN-PENKAL; FAGUNDES, em preparo).

substantivo ou do adjetivo, e em quarto lugar. Essa dissonância pode estar ligada à distribuição do léxico, que pode ser diferenciada nas três cidades e que será, posteriormente, objeto de análise mais refinada.

Assim, a pesquisa que começou para testar a hipótese de um maior conservadorismo em Irati vai revelando nuances de diferenças em diversos níveis linguísticos — morfonológico, léxico — e, provavelmente, com diferenças sociais. Podemos nos questionar com relação a essas últimas, levando em conta a relevância das variáveis sociais nas três localidades, evidenciada pela amostra. Nas três cidades, os fatores sociais — *sexo*, *idade* e *escolaridade* — foram considerados relevantes do ponto de vista estatístico. No entanto, a sua distribuição dá o que pensar: em Irati e Londrina, a *escolaridade* foi selecionada imediatamente após duas variáveis linguísticas. Em Pato Branco, foi a última de todas as variáveis e, consequentemente, depois de *sexo* e *idade*, selecionadas nas três cidades. Para uma melhor visualização, apresentamos abaixo o quadro da distribuição dos GFs.:

Quadro 1 – Variáveis selecionadas por cidades

Irati	**Londrina**	**Patro Branco**
1. Relação com o núcleo	1. Marcas precedentes	1. Posição linear
2. Posição linear	2. Classe gramatical	2. Relação com o núcleo
3. Escolaridade	3. Escolaridade	3. Animacidade
4. Grau do substantivo ou adjetivo	4. Processos morfológicos de formação de plural	4. Grau do substantivo ou adjetivo
5. Processos morfológicos de formação de plural	5. Sexo	5. Sexo
6. Sexo	6. Animacidade	6. Idade
7. Animacidade	7. Idade	7. Escolaridade
8. Idade	8. Relação com o núcleo	
9. Classe gramatical		

A partir dessa distribuição, podemos nos indagar qual é o papel da escolaridade na realização da regra de CN. Será que o fato de selecionar escolaridade entre os primeiros grupos de fatores, os mais relevantes, está ligado ao fato de a escola ser mais competente na implementação da variante mais padronizada? Ou será o fato de a CN ser

estigmatizadora dos falantes e as duas cidades mais sensíveis a essa marca social revelariam que o nível de escolaridade se reflete na aplicação da regra? Se o último caso fosse levado em conta, talvez se explique o fato de não ser selecionado em Pato Branco que, apesar de apresentar índices semelhantes aos de Irati e Londrina, nos três graus de escolaridade, teria um comportamento de concordância maior no geral. Essa hipótese é escudada pelos resultados de Dias, que apontam ser os italianos os que mais fazem concordância nominal nos predicativos e particípios passivos, quando comparados com os eslavos (Irati) e os açorianos (Florianópolis). Por que podemos afirmar isso? Ora, os colonizadores gaúcho-catarienenses do sudoeste do Paraná (onde se localiza Pato Branco) eram, em sua maioria, descendentes de imigrantes italianos que aportaram no RS sobretudo na segunda metade do século XIX. Os outros gaúcho-catarinenses que ajudaram a colonizar o Paraná, estariam dentro da perspectiva de Fernandes pois, junto com os italianos, foram os alemães os que mais aplicaram a regra de CN no sintagma nominal.

2.2 O que acontece quando juntamos resultados?

Como dispomos de um banco de dados constituído para tentar verificar se o português falado na região sul teria sofrido a influência de vários grupos migratórios (tanto internos quanto externos), a etapa seguinte seria, inevitavelmente, juntar os arquivos para fazer rodada geral que, se acredita, iria destacar diferenças dialetais. Foi o que fizemos, a partir de um arquivo em que constavam somente as variáveis comuns às três cidades: duas linguísticas, *posição em relação ao núcleo* e *animacidade* e as três sociais: *sexo*, *idade*, *escolaridade*, acrescida agora da *etnia*. Na rodada conjunta o programa considerou relevantes somente quatro: *posição em relação ao núcleo*; *sexo*, *escolaridade* e *cidades* (ou *etnia*). A partir dessa constatação, podemos levantar uma série de questões:

1. Uma das premissas da sociolinguística é a de que a variação vai ser fortemente condicionada por fatores sociais. Ora, na nossa amostra conjunta, é uma variável estrutural que se revela ser a mais importante na aplicação da regra, apontando para o fato de que a concordância nominal se constitui em variação inerente do português do

Brasil. Nesse sentido, a nossa hipótese de vocábulo fonológico reforça essa constatação e as variáveis sociais teriam um papel relativizado.

2. Embora tenha sido selecionado nas três cidades, o fator social *idade* desaparece quando se roda o conjunto dos dados. Isso eliminaria, em tese, toda a possibilidade de se falar em mudança em progresso. Essa constatação reforçaria a tese de variação inerente?

3. Se *idade* é descartada na rodada conjunta, *sexo* é a primeira das variáveis sociais selecionada (feminino 0,52 e masculino 0,47). Se mulheres fazem mais CN, de duas uma: ou a CN não é estigmatizada, o que não parecer ser o caso, inclusive pelos resultados de outras considerações a respeito da CN; ou se prova que as mulheres são, realmente, mais conservadoras. Contudo, embora sendo a primeira a ser selecionada, os pesos estão muito próximos do ponto neutro (0,50) e entre si (diferença e 5 pontos). Assim, somente o cruzamento dos dados poderia, talvez, elucidar a questão.

4. A *escolaridade* que, nas rodadas individualizadas das cidades, se mostrava como possível fator diferenciador entre Irati e Londrina de um lado, e Pato Branco, de outro, se neutraliza na rodada geral.

5. Rodando todos os dados juntos, o programa seleciona como relevante a variável *cidade*. Porém, como distinguir, nos resultados, o que é diferente em cada uma? Se, acima, apontamos uma neutralização da variável escolaridade, que era distinta em Pato Branco, nas rodadas individuais, no cômputo das cidades é Pato Branco que vai apresentar o maior índice de aplicação da regra de concordância 0,60, em comparação com a menor aplicação, 0,43, de Londrina. Irati que, em outros fenômenos se mostrava mais conservadora, na CN aparece com distribuição equitativa, 0,50, exatamente o ponto neutro. No entanto, pode se argumentar que esses números não são conclusivos, uma vez que se Pato Branco favorece a regra com 0,10 acima do ponto neutro, Londrina a desfavorece na proporção inversa. O que se tem de levar em consideração, nesse caso, é a distância entre o favorecimento e o desfavorecimento na aplicação da regra: temos 0,17 entre a menor e a maior aplicação da regra.

Podemos questionar se reunir dados, então, é a melhor solução. Parece que não, tendo em vista outras experiências com análises conjuntas, como a de Loregian-Penkal (2004), que demonstrou como esse tipo de análise pode mascarar a realidade dos fatos.

Na análise da alternância pronominal *tu/você* no Sul, empreendida por Loregian-Penkal (2004), ficou evidente que os resultados podem ficar mascarados quando se junta muitos dados e localidades na análise. A pesquisa foi feita em todas as localidades que integram o Banco VARSUL (Florianópolis; Ribeirão da Ilha; Lages; Chapecó e Blumenau, em Santa Catarina. Porto Alegre; Flores da Cunha; Panambi e São Borja, no Rio Grande do Sul. Os dados do Paraná não foram considerados para esta análise porque a ocorrência do pronome *tu* é praticamente inexistente na fala de paranaenses).

A análise não se restringiu à variação na comunidade visto que, nem sempre, a amostra é homogênea em sua constituição, pois pode acontecer de termos informantes que para determinado fenômeno sejam categóricos no tocante à realização da variável em estudo. Assim, a análise contemplou também a variação no indivíduo, conforme evidencia o Quadro 2 abaixo.

Quadro 2 – Reprodução da Tabela 07 de Loregian-Penkal (2004, p. 127): Usos dos pronomes *tu*/*você* por informantes

Localidade	**só TU**	**só VOCÊ**	**TU/VOCÊ**	**TU + [T+V]**	**VOCÊ + [T+V]**
Florianópolis	13	01	10 = 24	23	11
Porto Alegre	14	01	09 = 24	23	10
Ribeirão da Ilha	07	-	04 = 11	11	04
Chapecó	06	02	16 = 24	22	18
Blumenau	02	04	17 = 23	21	20
Lages	01	06	17 = 24	18	23
Flores da Cunha	13	-	10 = 23	23	10
Panambi	07	-	14 = 21	21	14
São Borja	14	01	06 = 21	20	07
TOTAL	77	15	103 = 195	180	118

Dos números apresentados, constata-se que 92 informantes se mostraram categóricos (77 usaram *só tu* e 15 *só você*). Note-se também a diferença entre esses números: há muito mais informantes categóricos no uso de *tu*, o que pode ser um indício da importância que esse pronome exerce na maioria das localidades analisadas.

Por outro lado, há um número significativo de 103 informantes que fazem uso da alternância *tu*/*você*. Se somados aos categóricos, temos o seguinte panorama: 180 falantes têm *tu* mais *tu*/*você* em sua gramática e 118 têm *você* mais *tu*/*você*. Como vemos, os falantes com *tu em* sua gramática continuam em maior número. Mas, mesmo assim uma análise[6] à parte dos indivíduos que têm ambos os pronomes se fez necessária. E essa análise foi efetuada pela pesquisadora.

A respeito de se ir além da análise na comunidade, Menon e Loregian-Penkal em 2002 já demonstraram que, se tratarmos só da variação na comunidade – só do todo – as diferenças podem se diluir e não se consegue dar uma explicação ao fenômeno da variação e ao da mudança, caso esta estiver ocorrendo. Para justificar a análise da variação no indivíduo, as autoras apresentaram os seguintes argumentos:

> Poderíamos afirmar que na região Sul do Brasil se alternam os pronomes *tu/ você* para representar a segunda pessoa do singular. Um olhar mais aprofundado, contudo, evidenciaria diferenças regionais não negligenciáveis: de um lado, Curitiba apresenta, categoricamente, o emprego de *você*. De outro, Porto Alegre, Florianópolis e Chapecó não têm informantes mulheres que usem categoricamente *só você*. Em Blumenau, tanto homens quanto mulheres preenchem as células de **só tu**, **só você** e **t+v**. Deparamo-nos, assim, com uma multiplicidade de distribuição que uma análise restrita à variação na comunidade mascararia. Como explicar que em Florianópolis há mais concordância canônica que em Porto Alegre se nesta há mais informantes usando *só tu*? Como saber qual o comportamento dos falantes que têm, na sua gramática, ambas as formas? (MENON; LOREGIAN-PENKAL, 2002, p. 161).

A conclusão a que chegam as autoras citadas é que é imprescindível analisar também o comportamento do indivíduo, averiguando, assim, o quanto o falante reflete o comportamento do grupo e vice-versa. Além disso, neste caso específico, uma análise restrita à comunidade provavelmente mascararia a forma como se encontram distribuídos os pronomes *tu/você* na amostra, assim como a concordância com o *tu* (regra variável também analisada por Loregian-Penkal (2004)). Soma-se a isso o fato de que, conforme aponta Guy (1980, p. 1), esse tipo de conhecimento da estrutura da variação parece ser indispensável para o entendimento dos processos históricos da mudança linguística e, também, para o estudo sincrônico da língua e seu uso social.

Dessa forma, a análise da variação no indivíduo efetuada por Loregian-Penkal (2004) demonstrou que é possível se considerar o todo, mas sem se esquecer das partes

[6] A análise sobre os indivíduos com ambos os pronomes é apresentada em detalhes no trabalho de LOREGIAN-PENKAL (2004).

fundamentais que o compõem. Mostrou ainda que esse tipo de refinamento da análise pode evitar conclusões errôneas a respeito de um determinado fenômeno em estudo.

Além disso, temos que levar em consideração um outro aspecto: a rodada geral permite constituir uma ortogonalidade que as amostras individualizadas podem não conter: uma amostra muito pequena pode não preencher todas as células de maneira equilibrada. Como a ortogonalidade é justamente o equilíbrio na distribuição dos dados nos diferentes grupos de fatores que compõem uma determinada amostra (que nem sempre é homogênea em sua constituição) o conjunto dos dados, rodados juntos, pode vir a dar à amostra geral a distribuição ótima para a análise estatística, mas nem sempre vai ser a melhor para explicar as (possíveis) especificidades locais. Nesse caso, o mais indicado seria fazer um refinamento da análise, considerando o indivíduo (no nosso caso, a cidade), obtendo, assim, uma descrição mais confiável para dar conta do comportamento da variável em estudo.

À guisa de conclusão

Se em Pato Branco a escolaridade não se mostrou tão relevante, por ser o último dos GFs. Selecionados mas, na rodada geral Pato Branco é a cidade que mais apresenta concordância, poderíamos concluir que a escolaridade faz alguma diferença? Nas duas cidades em que a escolaridade é selecionada como mais relevante, seria porque existe menos concordância por parte dos falantes, constituindo uma marca social e aí a escolaridade teria um papel a desempenhar? Lembremos que Londrina foi fundada e colonizada por migrantes mineiros e paulistas, conservando características dialetais marcadas nessas populações como discriminatórias, como a pronúncia do chamado **r** caipira, além da chamada falta de concordância, já apontada por Amadeu Amaral, no início do séc. XX, como uma das características marcantes desse dialeto. Em Irati, pode estar concorrendo para a escolaridade ser relevante o fato de boa parte da população ser de origem eslava. Ora, nas línguas eslavas, existe também um problema com a pronúncia do **r**, visto que elas só dispõem de um rótico, de pronúncia mais próxima ao do nosso tepe. Coincidência ou não, pode ser que o estigma da pronúncia do rótico se estenda à da concordância nominal em Irati, pelo fato de não fazer concordância estar

ligado à imagem das pessoas da área rural[7], normalmente de baixa ou nenhuma escolaridade, exatamente como aquelas ascendentes da população de Londrina.

Os nossos "achados" podem vir a ser úteis àqueles pesquisadores que vão analisar dados provenientes de mais de uma localidade, por exemplo, ou àqueles que pretendem descer às considerações sobre a variação no indivíduo *versus* variação na comunidade.

Referências

AMARAL, A. *O dialeto caipira:* gramática, vocabulário. 4. ed, São Paulo: Hucitec/ Brasília: INL, 1982 (reprod. facsimil da 2ª ed.; 1ª ed. 1920).

BANDEIRA, G. A. F. *O apagamento de se nas funções sujeito e objeto:* um estudo variacionista com dados do VARSUL do Paraná. Tese de doutorado. Curitiba, UFPR. 2007.

BISOL, L.; MENON, O. P. S.; TASCA, M. VARSUL, um banco de dados. In: VOTRE, S.; RONCARATI, C. (orgs.). *Anthony Julius Naro e a lingüística no Brasil:* uma homenagem acadêmica. Rio de Janeiro: 7 Letras/ FAPERJ. p.50-58. 2008.

DIAS, J. F. V. *A concordância de número nos predicativos e nos particípios passivos na fala da região Sul:* um estudo variacionista. Dissertação de Mestrado. Florianópolis: UFSC. 1996.

FAGUNDES, E. D. *As ocorrências do modo subjuntivo nas entrevistas do VARSUL no estado do Paraná e as possibilidades de variação com o modo indicativo.* Tese de doutorado. Curitiba, UFPR. 2007.

FERNANDES, M. *Concordância nominal na região Sul.* Dissertação de Mestrado. Florianópolis: UFSC. 1996.

GUY, G. R. Variation in the group and the individual: The case of final stop deletion. In Labov, W. ed., *Locating language in time and space.* New York: Academic Press: 1-36, 1980.

LOREGIAN-PENKAL, L. *(Re)análise da referência de segunda pessoa na fala da Região Sul.* Tese de doutorado. Curitiba, UFPR, 2004.

LOREGIAN-PENKAL, L. FAGUNDES, E.D. MENON, O. P. S. A. Análise da concordância

[7] Os imigrantes eslavos vieram sobretudo na condição de agricultores, como quaisquer outros imigrantes. No entanto, historicamente parece haver diferenças no domínio das técnicas e instrumentos agrícolas em uso pelos diferentes povos europeus, na época da sua imigração ao Brasil. Assim, os povos eslavos aqui chegados encontravam-se numa condição menos favorável, do ponto de vista do domínio de técnicas agrícolas, em relação aos outros imigrantes.

nominal em Irati e Pato Branco, Paraná. *Estudos Linguísticos.* Vol. 40, n.2, 2011.

MENON, O.P.S.; FAGUNDES, E.D.; LOREGIAN-PENKAL, L. O que fazer com grupos de fatores não selecionados? O caso da concordância nominal no Paraná. *Gragoatá*, Niterói, *29*: 147-160. 2010.

_____. *O meus filho:* a questão do vocábulo fonológico (em preparo).

_____. The VARSUL Database. Linguistik online 38, 2/2009. Disponível em http://www.linguistik-online.de/38_09/menonEtAl.html. Acesso em 10.03.2012.

SCHERRE, M. M. P. *Reanálise da concordância de número em português.* Rio de Janeiro, UFRJ. Tese de Doutorado, 1988.

Recebido em março de 2013.
Aceito em junho de 2013.