A importância da organização de uma base de dados para pesquisas atuais e futuras

Cláudia de Souza Cunha[1]

[1]Faculdade de Letras - Universidade Federal do Rio de Janeiro (UFRJ)
claudiascunha@ufrj.br

*Abstract**:** This article focus on linguistic database organization in Brazil and emphasizes studies of oral practise, regarding its theoretical and methodical evolution and its current perspectives. Two chronological phases can be distinguished: a phase that prioritizes a regional investigation – characterized by dialectal monographies dated from the first half of the twentieth century and by linguistics atlas published during following decades; and a second one that incorporates principles of quantitative sociolinguistics, originating new "corpora" – as NURC, CENSO and VARSUL – horizontally and vertically stratified. Finally, we present the "Brazilian Linguistic Atlas Project", which has provided a stratified database, based on geosociolinguistics, without precedent either in dimension or in embracement of linguistic phenomena.*

*Keywords**:** geolinguistc; sociolinguistic; linguistic database.*

Resumo: Este artigo enfoca a organização de bancos de dados lingüísticos no Brasil voltados para o estudo da oralidade, considerando sua evolução teórico-metodológica e as perspectivas atuais. Destacam-se duas fases: a que prioriza a investigação diatópica – marcada pelas monografias dialetais da primeira metade do século XX e pelos Atlas lingüísticos publicados nas décadas seguintes; e a fase que incorpora os preceitos da sociolingüística variacionista, originando novos "corpora" – como NURC, CENSO e VARSUL – estratificados horizontal e verticalmente. Por fim, apresenta-se o "Projeto Atlas Lingüístico do Brasil", que vem levantando um banco de dados estratificado, de orientação geo-sociolingüística, inédito quanto à dimensão e à abrangência dos fenômenos lingüísticos enfocados.

Palavras-chave: geolingüística; sociolingüística; base de dados lingüísticos.

1. Introdução

Quando me foi proposto abordar este tema no Simpósio "Pesquisas dialetológicas e geolingüísticas no Brasil: um olhar histórico-metodológico", junto a outros membros da equipe do Projeto ALiB neste *53º Seminário do GEL*, veio-me à lembrança um texto de Sônia Costa, pesquisadora da UFBA, escrito a propósito de uma Semana de Letras realizada na Universidade do Sudoeste da Bahia (UESB), em 2003, e que tem por título "A lingüística e os estudos da linguagem rumo ao século XXI". Tomando por base as sugestões de Castilho 2001 e os depoimentos de dezoito lingüistas brasileiros coletados por Xavier e Cortez 2003, a autora lista 14 tarefas, algumas já iniciadas e outras por iniciar, que, a seu ver, "parecem importantes e promissoras para o futuro próximo da Lingüística". Em síntese, cabe à lingüística:

1. aprofundar as descrições no âmbito sincrônico e diacrônico da língua portuguesa em todas as suas variedades e também das línguas indígenas, constituindo bancos de dados de grande porte e produzindo obras de referência;

2. aprofundar e alargar os limites das abordagens teóricas, explorando, por exemplo, o cognitivismo – que compreende questionamentos acerca do desenvolvimento lingüístico e sua relação com a cognição;

2. investir na interface com a computação;

3. contribuir para o ensino de língua portuguesa e, mais amplamente,

4. fornecer assessoria ao Estado no que se refere às questões de ordem lingüística, atentando inclusive para as possibilidades que se abrem para os profissionais de Letras com o Mercosul.

Quando postas as tarefas da lingüística assim em conjunto, sobressai a necessidade de se organizarem bases de dados para a pesquisa, de modo que, no tocante especialmente à dialectologia, se possam inventariar, sistematizar e interpretar as variantes da língua portuguesa observando a distributividade – espacial, cronológica, sociocultural etc – dos traços lingüísticos depreendidos.

2. A oralidade como objeto de estudo

Em termos históricos, Rossi (*Enciclopédia Mirador*, Verbete "dialectologia", p.3298) relembra que

> *O conhecimento empírico, às vezes parcial ou intuitivamente sistematizado, da "diversidade na unidade lingüística", documenta-se na cultura ocidental pelo menos desde a Grécia antiga, onde se reconhecia a existência de quatro dialetos, o eólico, o dórico, o jônico e o ático.*

Mas é no século XIX que os estudiosos voltam os olhos para a modalidade oral, por um desses atalhos a que a ciência agradece. Os neogramáticos, deparando-se cada vez mais com irregularidades e exceções às leis históricas formuladas, não cogitaram rever a convicção teórica de que as mudanças fonéticas se subordinavam a uma regularidade praticamente absoluta. Antes atribuíram as irregularidades à natureza das línguas investigadas, por serem, segundo Nelson Rossi, "resultado daquilo a que hoje se chamaria contatos e interferências (na época, 'misturas', 'contaminações', ou termos equivalentes)." Raciocinava-se então que uma língua isenta de "misturas" em seu desenvolvimento histórico estaria mais apta a comprovar a regularidade das leis fonéticas. Manteve-se o método e mudou-se o *corpus*. Assim, sob a influência evidente da ideologia romântica do século XIX, criou-se o conceito da "língua pura". Para encontrá-la "impunha-se recorrer às manifestações lingüísticas concretas do povo, principalmente na comunicação oral".

E como conclui Rossi

> *A conseqüência foi uma ampla e intensa valorização, como objeto primário de estudo, das variedades lingüísticas usadas por falantes iletrados, principalmente aqueles que pertencessem a comunidades isoladas, pouco ou supostamente nada sujeitas a influências de outras comunidades.*

Engendrada a idéia do *corpus* ideal, a dialectologia configura-se, paulatinamente, como área de estudo, definindo metas e aprimorando seus métodos – principiando a coleta de dados dialetais através de inquéritos realizados por correspondência (na França, em 1807), até consolidar-se, nesta 1ª fase, com a publicação do Atlas Lingüístico da França (cujo 1º volume data de 1902).

Sendo este Simpósio dedicado às pesquisas dialectológicas e geolingüísticas no Brasil, permitam-me cruzar o Atlântico e rememorar, brevemente, o que, ainda no raiar do século passado, produziu-se a respeito dos nossos falares.

3 – Os *Corpora* de língua falada no Brasil

A 1ª metade do século XX é marcada pelo aparecimento de estudos precursores, de cunho monográfico, com as relevantes contribuições de Amadeu Amaral (*O dialeto caipira*, de 1920), Antenor Nascentes (*O linguajar carioca*, de 1922, reeditado em 1953) e Mário Marroquim (*A língua no nordeste – Alagoas e Pernambuco*, de 1934). Os três se propõem a descrever "a língua do povo", abordando a fonética, a morfossintaxe e o vocabulário. Mas é em 1963, com a publicação do *Atlas Prévio dos Falares Baianos (APFB)*, que se apresenta, pela 1ª vez, um *corpus* oral, representativo de uma região do Brasil, sistematizado segundo os parâmetros da dialectologia tradicional: a) para a coleta dos dados, aplicou-se um questionário; b) para a seleção das localidades, observaram-se critérios como antigüidade e identidade histórico-cultural relevante; c) para a seleção do informante, buscou-se o falante iletrado, não raro de idade avançada, tido por seus pares como representante genuíno de sua comunidade.

Além do APFB, oito atlas regionais foram publicados, documentando sob a forma de cartas lingüísticas a fala popular de 9 estados brasileiros: o *Esboço de um Atlas Lingüístico de Minas Gerais* (1977), o *Atlas Lingüístico da Paraíba* (1984), o *Atlas Lingüístico de Sergipe* (1987), o *Atlas Lingüístico do Paraná* (1994), o 2º Volume do *Atlas Lingüístico de Sergipe* (2002), o *Atlas Lingüístico-Etnográfico da Região Sul* (2002), o *Atlas Lingüístico Sonoro do Pará* (2004) e o *Atlas Lingüístico do Amazonas* (2004).

Representantes de gerações distintas, os atlas se diferenciam metodologicamente conforme o perfil do informante, o tipo de questionário e o tipo de carta lingüística. Consideremos por agora o 'perfil do informante'. Se os primeiros atlas (o da Bahia, o de Minas Gerais e o da Paraíba) observavam apenas a escolaridade para a seleção, os atlas mais recentes (o do Pará e o do Amazonas) incorporam sistematicamente os fatores sexo e faixa etária, distribuindo-os tal como preconiza a sociolingüística variacionista.

No método laboviano, a amostra deve ser estratificada, de modo a que se tenham dados comparáveis entre si; e aleatória, de forma a garantir que os informantes – que não devem ser escolhidos segundo características individuais – sejam representantes da realidade lingüística da população escolhida para a pesquisa. Para tanto, divide-se a população em células compostas, cada uma, de indivíduos com as mesmas características sociais, procedendo-se, posteriormente, para preencher cada casa, a uma seleção aleatória, conforme exemplificam, didaticamente, Mollica e Braga (2003:121).

> *Se for escolhida como objeto de pesquisa apenas a variável social sexo, pode-se ter, numa casa, 5 homens e 5 mulheres, e a amostra poderá ser teoricamente de 10 indivíduos. Se acrescentarmos a variável classe social, por exemplo, e dividirmos essa variável em três fatores correspondendo às classes alta, intermediária e baixa, já teremos de ter as seguintes casas: 5 homens da classe alta;5 homens da classe*

intermediária; 5 homens da classe baixa; 5 mulheres da classe alta; 5 mulheres da classe intermediária; 5 mulheres da classe baixa.

Entre as décadas de 70 e 90 criaram-se importantes acervos de base sociolingüística, destacando-se o do NURC (que abarca 5 capitais do país), o do PEUL (que enfoca a região metropolitana do Rio de Janeiro) e o do VARSUL (que recobre os três estados da Região sul). Retomo aqui o perfil de cada um deles.

O PEUL[1] - Programa de Estudos sobre o Uso da Língua – é um grupo de pesquisas inter-institucional que reúne professores da UFRJ (onde o grupo está sediado), da UFF e da UNB, voltados para o estudo do binômio variação e mudança e da interface língua sociedade. Conjugando diferentes perspectivas teóricas (funcionalista, gerativista, gramaticalização, teorias do discurso) aos pressupostos teórico-metodológicos da Sociolingüística Laboviana o grupo tem procurado contribuir, ao longo da sua existência, para a compreensão da realidade lingüística brasileira, através de trabalhos que enfocam sua complexidade sócio-geográfica.

A principal amostra que integra o acervo do projeto é a Amostra CENSO – 80 **(**Censo da Variação lingüística do Rio de Janeiro). A amostra é composta de gravações de 64 falantes da cidade do Rio de Rio de Janeiro, realizadas entre 1980 e 1984. Constituída com o objetivo de propiciar estudos de processos de variação e mudança na variedade carioca "não-culta", registra a fala de indivíduos de ambos os sexos e três níveis de escolaridade: 1º segmento do Ensino Fundamental (EF-1), 2º segmento do Ensino Fundamental (EF-2) e Ensino Médio (EM):

Tabela 1. Distribuição da Amostra CENSO – 80

Idade	7-14 anos		15-25 anos		26-49 anos.		+ de 50 anos	
Sexo	H	M	H	M	H	M	H	M
EF-1	4	4	3	3	2	2	5	4
EF-2	4	4	3	2	3	3	3	2
EM	X	X	2	3	2	3	1	2
TOTAL	8	8	8	8	7	8	9	8

O Projeto VARSUL[2] (Variação Lingüística Urbana na Região Sul) foi constituído oficialmente em 1990 e visa à instalação de um Banco de Dados lingüísticos a partir da documentação do português falado nas áreas urbanas lingüisticamente representativas dos estados do Paraná, Santa Catarina e Rio Grande do Sul. Sua realização fica a cargo de uma equipe multi-institucional, da qual participam professores-pesquisadores vinculados a quatro instituições: Universidade Federal do Paraná, Universidade Federal de Santa Catarina, Universidade Federal do Rio Grande do Sul e a Pontifícia Universidade Católica do Rio Grande do Sul.

Trata-se de um trabalho realizado dentro dos postulados da Sociolingüística Variacionista, que pretende fornecer subsídios para estudos de variação lingüística da região sul do Brasil. A coleta de dados foi iniciada em 1990 e se estende até hoje. O Banco de dados VARSUL contém amostras representativas da fala de habitantes de 12 cidades, quatro em cada estado da Região Sul, num total de 96 entrevistas por estado, e 288 no acervo do Banco de Dados VARSUL. Trata-se de um trabalho realizado, igualmente, dentro dos postulados da sociolingüística variacionista, que pretende fornecer subsídios para estudos da variação lingüística da região.

Os informantes estão distribuídos por: a) sexo (homem e mulher); b) idade (25 a 50 anos e mais de 50 anos); c) nível de instrução (até 5, até 8/9 e até 11/12 anos de escolaridade); d) variedades lingüísticas: capitais e grupos étnicos ou sociolingüísticos culturalmente representativos de cada um dos estados.

Tabela 2. Distribuição da Amostra VARSUL

Idade	25-50 anos		Mais de 50 anos	
Sexo	H	M	H	M
Escolaridade: 5 anos	24	24	24	24
Escolaridade: 8/9 anos	24	24	24	24
Escolaridade:11/12 anos	24	24	24	24
TOTAL	72	72	72	72

O Projeto NURC[3] – Projeto de Estudo da Norma Lingüística Urbana Culta, vinculado ao "Proyeto de Estudio Coordinado de la Norma Lingüística Culta de las Principales Ciudades de Iberoamérica y de Península Ibérica", instalou-se em 1969 durante o III Instituto Interamericano de Lingüística, promovido em São Paulo. Por sugestão do Professor Nelson Rossi, sua abrangência estendeu-se às cinco principais capitais com mais de um milhão de habitantes: Recife, Salvador, Rio de Janeiro, São Paulo e Porto Alegre. O projeto previa três etapas: gravações, transcrição e análise do *corpus*, conforme um Guia-Questionário. Inicialmente, eram previstas 400 horas de gravação, selecionando-se 600 informantes (300 do sexo masculino e 300 do sexo feminino) com nível superior de escolaridade, nascidos na cidade sob estudo ou residentes aí desde os cinco anos de idade, filhos de nativos de língua portuguesa, de preferência nascidos na cidade sob pesquisa. Os informantes foram distribuídos em três faixas etárias: 1) de 25 a 35 anos de idade; 2) de 36 a 55 anos de idade; 3) mais de 56 anos de idade. Quanto à natureza, as gravações foram divididas em quatro tipos: Gravações secretas de um diálogo espontâneo (GS); Diálogo entre dois informantes (D2); Diálogo entre o informante e o documentador (DID); Elocuções Formais (EF). Em 1985, durante a XIII Reunião Nacional do Projeto NURC, realizada em Campinas, decidiu-se que as cidades intercambiariam 18 entrevistas de seu acervo com as demais cidades. Esse acervo constituiu-se o que se convencionou chamar *corpus compartilhado*.

Tabela 3. Distribuição da Acervo NURC

CIDADES PESQUISADAS	NÚMERO DE ENTREVISTAS	NÚMERO DE INFORMANTES	HORAS DE GRAVAÇÃO
RECIFE	363	461	307 horas e 20 minutos
SALVADOR	357	456	304 horas
RIO DE JANEIRO	394	493	328 horas e 40 minutos
SÃO PAULO	381	474	316 horas
PORTO ALEGRE	375	472	413 horas e 40 minutos
TOTAL	1870	2356	1669 horas e 40 minutos

As bases de dados apresentadas por estes acervos permitiram a verticalização do estudo da variação lingüística, sem no entanto perder-se – nas amostras NURC e PEUL

– a dimensão horizontal, garantida pela extensão regional de seus *corpora*, conforme mostra o mapa constante do Anexo 1.

4. O *corpus* ALIB

Chegamos ao século XXI com o seguinte quadro de registro da oralidade:

Há 8 atlas lingüísticos que registram a fala popular de 9 estados do Brasil, sendo que dois deles (o do Pará e o do Amazonas) contêm amostra estratificada, considerando os parâmetros sexo e faixa etária; tem-se o registro estratificado da fala culta de 5 capitais brasileiras (acervo NURC) e de 12 cidades (incluídas as capitais) dos estados do sul (amostra VARSUL); e há uma série de bancos de dados estratificados, de recorte mais específico, abrangendo uma única região (citem-se, por exemplo, a Amostra CENSO, o *corpus* APERJ, o Projeto ALIP – Amostra Lingüística do Interior Paulista).

Salvaguardados os méritos e a relevância desses *corpora* para o conhecimento do português brasileiro, se buscarmos, a partir de seus dados, tecer generalizações descritivas que aliem amplitude geográfica, variedades de perfis sócio-culturais e situações elocucionais, deparamo-nos com um entrave: falta-lhes intercomparabilidade.

Neste sentido, o Projeto ALiB (Projeto Atlas Lingüístico do Brasil) vem levantando um banco de dados estratificado, inédito quanto a dois aspectos. Quanto à dimensão, pois registrar-se-á a fala estratificada de 1.100 informantes, num total de 250 municípios de todo o Brasil, conforme se vê no mapa (Anexo 2) que ilustra a rede de pontos; e quanto à abrangência dos fenômenos lingüísticos enfocados, visto que fazem parte do questionário questões de cunho metalingüístico, pragmático e prosódico, bem como o registro de discursos semi-dirigidos, totalizando 436 questões.

O *corpus* ALiB, reunirá, em suma, as principais características dos grandes acervos sonoros disponíveis hoje:

Tabela 4. Comparação em entre os corpora NURC, PEUL, VARSUL e ALiB

	Fala		Zona		Elocução		Amostra	
	culta	não-culta	urbana capitais	urbana demais cidades	espontânea	Semi-dirigida	estratificada	Não estratificada
NURC	X		X		X		X	
PEUL		X	X		X		X	
VARSUL	X		X	X	X		X	
ALiB	X	X	X	X	X	X	X	

5 – Perspectivas de análise do *Corpus* ALiB

Um banco de dados amplo, constituído por critérios homogêneos, presta-se ao mapeamento e a uma descrição mais segura da variação lingüística no aspecto diatópico. O banco de dados fornecido pelos Atlas Lingüísticos é capaz de apontar tendências comportamentais dos fenômenos variáveis, como a carta 37 do ALERS, (constante do Anexo 3), que dialoga com as observações de Callou e Leite (2002, 47-48) acerca do comportamento da lateral posvocálica:

> *O material examinado confirma um fato público e notório: a vocalização da lateral é geral em todo o país. Não pode ser considerada um traço de fala popular restrito a algumas áreas, como Ceará e Rio de Janeiro, como se poderia pensar, a julgar pelos textos clássicos.Comparando a fala culta de*

> *São Paulo, Rio de Janeiro, Salvador, Recife e Porto Alegre, é fácil perceber que Porto Alegre se distingue das demais, sendo a realização alveolar/velar praticamente exclusiva dessa capital. Além disso, a pronúncia vocalizada concorre em Porto Alegre com a pronúncia velar/alveolar, com predomínio da primeira entre os jovens, o que indicaria uma mudança em progresso, mas apenas para os homens. As mulheres apresentam outra configuração de uso, em que se igualam jovens e idosos em contraste com os adultos, o que corresponde a uma variação estável.*

Atlas de última geração, como o ALAM, oferecem um painel que pode chegar a contemplar tendências relativas a três variáveis: região, faixa etária e sexo, como se vê na carta 43 (veja-se o Anexo 4), que enfoca 3 fenômenos fonéticos distintos: a realização da vibrante em coda silábica e a alternância e a realização do morfema indicador de gerúndio (-ndo).

6. Comentários finais

A base de dados oferecida pelo Atlas do Brasil, poderá, dentre muitas aplicações, fornecer subsídios para o conhecimento da prosódia regional (cuja descrição se restringe ainda às capitais do *corpus* NURC – Cunha 2000) e fomentar a discussão, por meio do registro da fala popular em todo o país, de questões controversas, como as origens do português brasileiro, a exemplo das reflexões de Serra 2003, baseadas em cartas do APFB.

Notas

[1] Informações obtidas no site oficial do Projeto: www.letras.ufrj.br/~peul

[2] Informações obtidas no site oficial do projeto: http://www.cce.ufsc.br/~varsul

[3] Informações obtidas em sites oficiais do Projeto: http://www.fflch.usp.br/dlcv/nurc e http://www.letras.ufrj.br/nurc/

Referências

CALLOU, Dinah e LEITE, Yonne. *Como falam os brasileiros*. Rio de Janeiro, Zahar, 2002.

CASTILHO, A. Políticas lingüísticas no Brasil: o caso do português brasileiro. In: *Lexis* XXV. 1 y 2. Lima: Departamento de Humanidades / Pontificia Universidad Católica del Perú, pp. 271-297, 2001.

COSTA, Sônia Bastos Borba. "A lingüística e os estudos da linguagem rumo ao século XXI". Texto apresentado na Semana de Letras da Universidade do Sudoeste da Bahia. Vitória da Conquista - UESB, 16.10.2003. Publicado na web: http://www.prohpor.ufba.br/alinguis.html

CUNHA, Cláudia. Entoação regional no português do Brasil. Tese de Doutoramento. Faculdade de Letras, UFRJ, 2000.

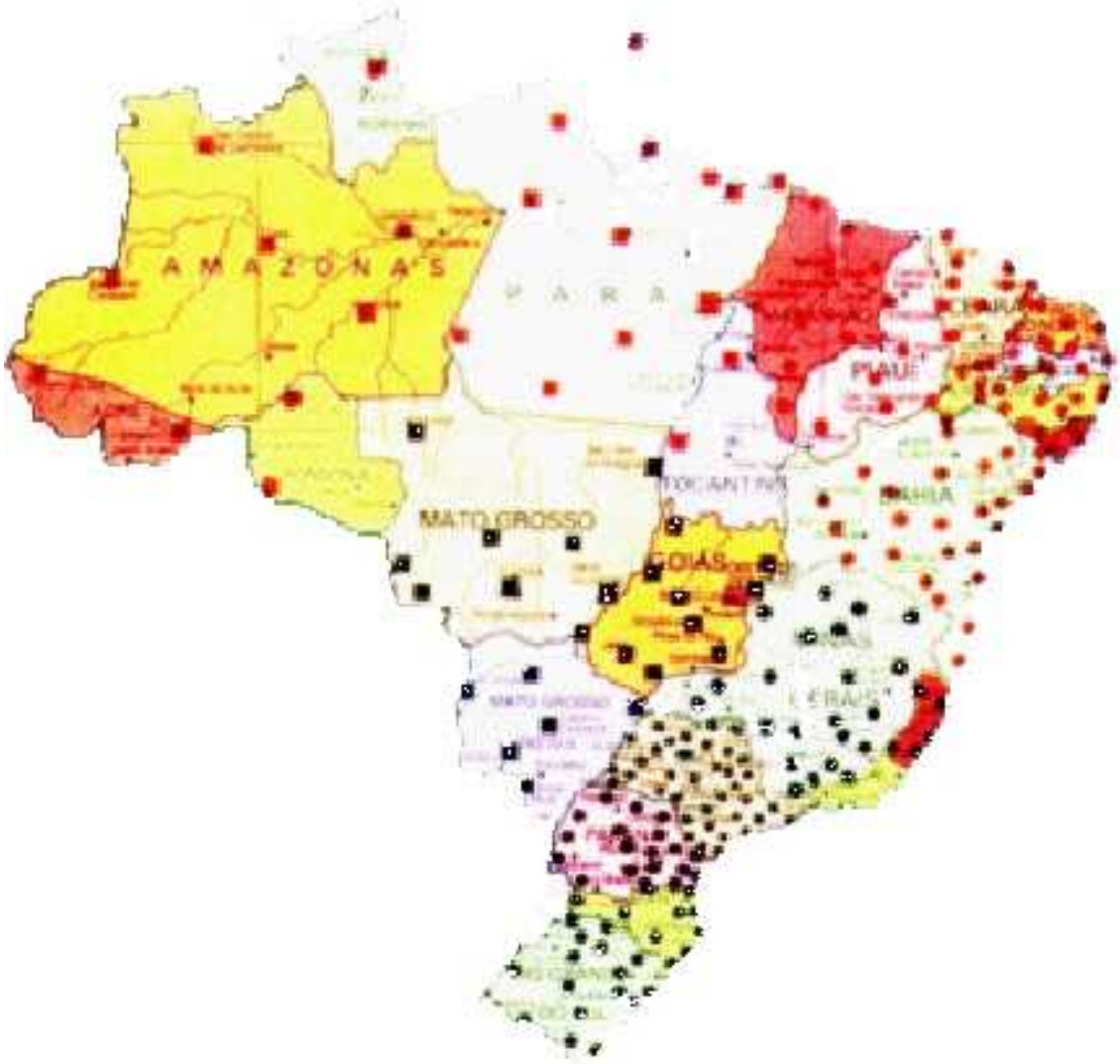
ROSSI, Nelson. "Verbete Dialectologia". *Enciclopédia Mirador*. p.3298-3303.

SERRA, Carolina Ribeiro. "Origens do português brasileiro". FL, UFRJ, 2003.

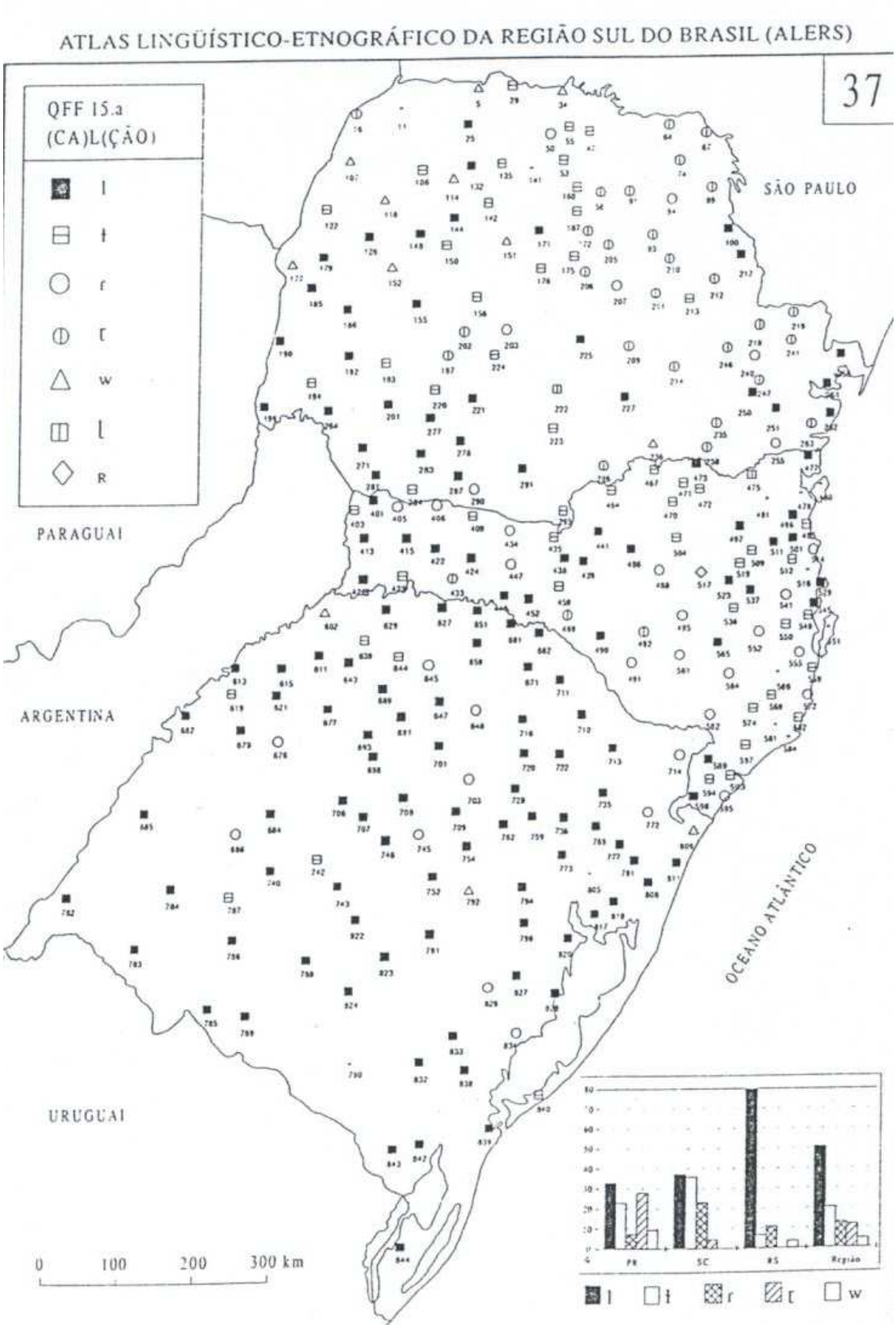
XAVIER, Antonio Carlos & Cortez, Suzana (orgs.). *Conversas com lingüistas*: *virtudes e controvérsias da Lingüística*. São Paulo: Parábola, 2003.

Anexo 1

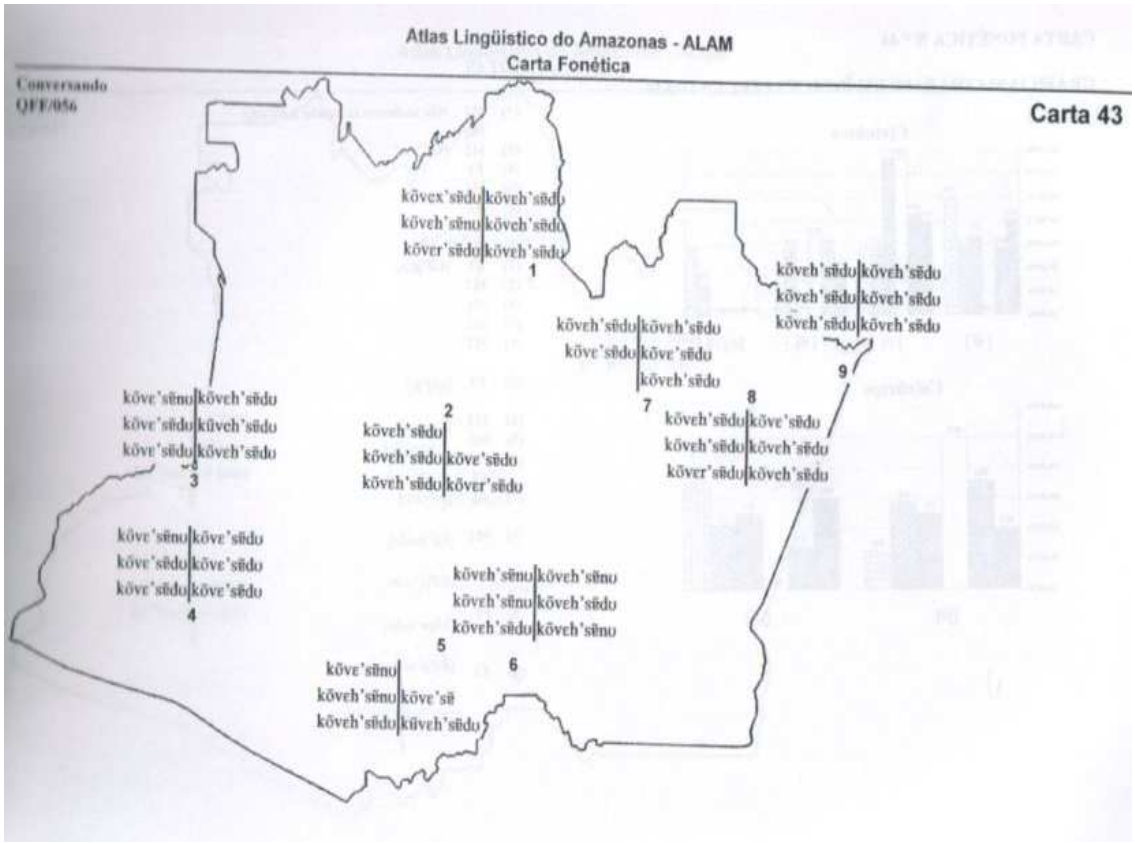
Anexo 2



Anexo 3


ATLAS LINGÜÍSTICO-ETNOGRÁFICO DA REGIÃO SUL DO BRASIL (ALERS)
37
QFF 15.a
(CA)L(ÇÃO)
l
t
r
[
w
L
R
SÃO PAULO
PARAGUAI
ARGENTINA
URUGUAI
OCEANO ATLÂNTICO
0
100
200
300 km
PR
SC
RS
Região
l
t
r
[
w

Anexo 4


Atlas Lingüístico do Amazonas - ALAM
Carta Fonética
Conversando
QFF/056
Carta 43
kõvex'sẽdu kõveh'sẽdu
kõveh'sẽnu kõveh'sẽdu
kõver'sẽdu kõveh'sẽdu
1
kõveh'sẽdu kõveh'sẽdu
kõveh'sẽdu kõveh'sẽdu
kõveh'sẽdu kõveh'sẽdu
9
kõveh'sẽdu kõveh'sẽdu
kõve'sẽdu kõve'sẽdu
kõveh'sẽdu
7
8
kõveh'sẽdu kõve'sẽdu
kõveh'sẽdu kõveh'sẽdu
kõver'sẽdu kõveh'sẽdu
kõve'sẽnu kõveh'sẽdu
kõve'sẽdu kũveh'sẽdu
kõve'sẽdu kõveh'sẽdu
3
2
kõveh'sẽdu
kõveh'sẽdu kõve'sẽdu
kõveh'sẽdu kõver'sẽdu
kõve'sẽnu kõve'sẽdu
kõve'sẽdu kõve'sẽdu
kõve'sẽdu kõve'sẽdu
4
kõveh'sẽnu kõveh'sẽnu
kõveh'sẽnu kõveh'sẽdu
kõveh'sẽdu kõveh'sẽnu
5
6
kõve'sẽnu
kõveh'sẽnu kõve'sẽ
kõveh'sẽdu kũveh'sẽdu

CARTA FONÉTICA N ° 43

GRÁFICO(S) COM BASE EM ÍNDICES PERCENTUAIS