Cadernos de Tradução

Instituto de Letras

Nº 1 – Janeiro de 1998

VARBRUL: ANÁLISE AVANÇADA[1]

Gregory R. Guy
York University, North York (Canadá)
Tradução: Ana Maria Stahl Zilles

0. Introdução. Este texto baseia-se em uma oficina ministrada durante o XVI NWAV[2]. Seu objetivo é oferecer ao usuário do Varbrul uma visão geral, de caráter prático, de alguns dos métodos e problemas associados à análise quantitativa (de regra variável) de dados lingüísticos pelo programa desenvolvido por Sankoff e Rousseau. Os aspectos tratados aqui não devem ser vistos como notáveis descobertas originais do autor; ao contrário, a maioria desses tópicos faz parte das técnicas tradicionais de análise estatística. Mas, entre lingüistas, em geral, essas informações costumam passar de boca em boca, uma versão acadêmica da tradição pré-letrada da história oral. Isso constitui um obstáculo à rápida disseminação de uma técnica essencial a uma comunidade de estudiosos interessados. A informação tem que estar à disposição numa versão escrita acessível aos lingüistas. É isso que este texto tenta fazer. Eu o escrevo não como um desbravador ou inovador, mas como um mero compilador das histórias contadas por aqueles que sabem. Quaisquer falhas de interpretação são, é claro, de minha responsabilidade.

O adjetivo "avançada", empregado no título deste texto, requer esclarecimento. O presente trabalho dirige-se àqueles que leram a respeito e talvez já tenham usado a metodologia Varbrul, e que gostariam de saber mais sobre como empregá-la melhor. Não é tão avançado a ponto de beneficiar um pesquisador experimentado na área, mas vai além do nível da mera introdução ao assunto. Assim, eu pressuponho pelo menos um conhecimento mínimo dos objetivos, técnicas e terminologia da análise Varbrul. Uma revisão muito breve desse conteúdo é fornecida abaixo, mas o leitor que não esteja minimamente familiarizado é aconselhado a revisar a literatura sobre

[1] N.T. Agradeço ao Prof. Gregory R. Guy a autorização para traduzir este texto, originalmente publicado em Ferrara, K. et al. (eds.) **Linguistic change and contact**. (Texas Linguistic Forum, vol.30) Austin: Univ. of Texas, Department of Linguistics, 1988. Agradeço, também, ao Prof. Álvaro Vigo, do Depto. de Estatística do Instituto de Matemática da UFRGS, pela leitura da tradução e por suas valiosas sugestões.

[2] N.T. O autor refere-se à XVI edição do encontro anual intitulado New Ways of Analyzing Variation.

o assunto como, por exemplo, Cedergren (1973), Cedergren & Sankoff (1974), Guy (1975), Sankoff (1975), Rousseau & Sankoff (1978), Sankoff & Labov (1979), Rousseau (1983).

1. Por que "regras variáveis"? A análise de regra variável foi desenvolvida na lingüística como uma forma de dar conta da variação estruturada, governada por regras, no uso da língua. Trata-se da variação lingüística que regularmente mostra taxas mais altas ou mais baixas de ocorrência em determinados ambientes; ou que regularmente predomina em determinados grupos sociais ou estilos de fala.

Um exemplo é o conhecido caso do /r/ posvocálico na cidade de Nova Iorque. Esse dialeto, originalmente "sem-r", está começando a importar uma pronúncia "com-r" como uma norma de prestígio dos dialetos circundantes, por vezes denominados de *General American*[3]. O uso de realizações consonantais de /r/ posvocálico em Nova Iorque é nitidamente estratificado por classe social, ajustado em cada situação de acordo com o estilo de fala e condicionado por aspectos do contexto lingüístico como acento e segmento seguinte. Esses padrões são altamente sistemáticos em todos os grupos sociais, conforme ilustrado na Figura 1, extraída de Labov (1966).

Não se trata de "variação aleatória" ou "livre", mas do que Weinreich, Labov & Herzog denominam de "heterogeneidade ordenada" (1968:100). Há claramente algumas "generalizações significativas do ponto de vista lingüístico [e social]" a serem capturadas aqui. Os grupos de classes mais altas usam mais /r/ em todas as ocasiões, e todos usam mais /r/ em seus estilos mais formais. Uma análise descritiva adequada dessa variação deveria definir claramente a natureza e a extensão de cada um desses efeitos condicionadores, permitindo "predizer" probabilisticamente a taxa aproximada de uso de /r/ por um certo indivíduo, dadas as informações sobre as características sociais da pessoa, da situação social e do contexto lingüístico.

[3]N.T. A expressão *General American* designa o dialeto mais geral dos EEUU, que inclui o Oeste, o Meio-Oeste (exceto a área dos Grandes Lagos) e parte do Leste, parte da Pensilvânia e de Nova Jérsei; distingue-se dos dialetos do Sul, de Nova Iorque, da Nova Inglaterra e dos Grandes Lagos.

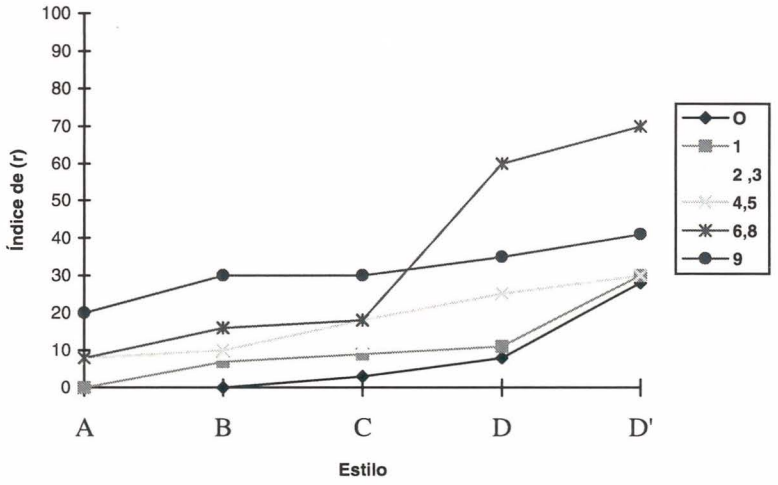

Figura 1 - Estratificação social de (r) na cidade de Nova Iorque (cf. Labov, 1966:240)

O objetivo da análise de regra variável é precisamente esse. Lidando com fenômenos variáveis, e não categóricos[4], ela é necessariamente quantitativa. Qualquer generalização que possamos formular sobre a Figura 1 não poderia ser refutada por um contra-exemplo isolado, mas apenas por um teste estatístico realizado através de um conjunto de dados. Assim, a análise vai envolver a contagem das ocorrências da variável, a descrição de tendências e da extensão da variabilidade, bem como das restrições ou fatores que a influenciam, mediante métodos estatísticos.

Um problema analítico surge aqui, do fato de que não podemos fazer experimentos controlados que isolem cada um dos fatores a fim de testar seus efeitos sobre a variável em questão. Cada ocorrência de uma variável vem inserida em um enunciado e em um contexto social, que poderia abarcar um grande número de fatores que influenciam a escolha do falante, feita a partir de uma gama de alternativas. Portanto, a análise tem que ser necessariamente

[4]N.T. Mantive os adjetivos variável e categórico por serem de largo uso na lingüística. Na literatura estatística, empregam-se os termos aleatório e determinístico, respectivamente. Fenômenos aleatórios são regidos por leis probabilísticas, enquanto fenômenos determinísticos são, usualmente, regidos por leis físicas ou químicas. Estes, no contexto lingüístico, resultariam da aplicação de regras obrigatórias ou categóricas.

multivariada; em outras palavras, ela é uma tentativa de modelar os dados como uma função de várias forças simultâneas, interseccionadas e independentes, que podem estar atuando em diferentes direções. De fato, um dos produtos da análise é uma medida numérica do peso e da "direção" (favorável ou desfavorável) de cada força.

Tal análise deve basear-se em um grande *corpus* de observações da variável em muitos contextos diferentes e envolve a parcialização dos diversos efeitos, controlando-se um de cada vez. A Figura 1 ilustra como isso é feito. Discernimos os efeitos do estilo traçando a linha correspondente a cada grupo social. Seguindo a linha da esquerda para a direita, a inclinação é sempre ascendente. Assim, concluímos que, independentemente da classe social de um falante, estilos mais formais sempre evocam mais casos de /r/ pronunciado. Um contra-exemplo seria uma linha descendente: de fato, não ocorre nenhuma.

Da mesma forma, identificamos o efeito da classe controlando o estilo. Examinando a coluna correpondente ao estilo B, por exemplo, encontramos diferentes grupos socioeconômicos claramente estratificados, do status mais alto ao mais baixo. Quando observamos que o mesmo padrão também é verdadeiro nos estilos C e D, começamos a generalizar dizendo que o *status* se correlaciona com a pronúncia do /r/. Mas há leves discrepâncias nos pontos estilísticos extremos. Se temos ou não, num tal caso, violações significativas do padrão é uma questão que poderá ser respondida por meio de testes estatísticos.

O ponto importante aqui é que somente essas comparações controladas permitem que façamos generalizações válidas. Comparar os falantes da classe 9 no estilo B com os falantes da classe 1 no estilo D não revela nada sobre estilo ou classe. Ao contrário, uma ou outra das dimensões deve ser constante.

Para alguns leitores pode ser útil revisar a terminologia da análise de regra variável introduzida por Labov (1969) e Cedergren & Sankoff (1974). Uma "regra variável" é uma regra de reescritura sensível ao contexto que relaciona um par de variantes como x —> <y>, de modo que, quando a regra se aplica, ocorre "y" e, quando não se aplica, ocorre "x". Cada uma das dimensões analíticas do contexto que determina se a regra se aplica ou não é chamada de "grupo de fatores". No exemplo, ESTILO e CLASSE SOCIOECONÔMICA são os dois grupos de fatores. Cada grupo é formado de um conjunto de

"fatores" discretos; por exemplo, o grupo de fatores estilo compreende cinco fatores, A, B, C, D e D'. Cada combinação única de fatores descreve uma "célula"; por exemplo, [SEC 9, estilo A][5] é uma das células da Figura 1. Na análise quantitativa, cada fator recebe um "peso" ("valor do fator"ou "probabilidade") — um número entre zero e um que caracteriza o efeito deste fator sobre a regra variável em questão (quanto mais alto o número, maior a chance de a regra se aplicar quando este fator estiver presente no contexto). E, finalmente, um conjunto total de dados é caracterizado por uma "probabilidade de *input*[6]" ("p_0"), que é, de fato, uma medida global do índice de aplicação da regra. Uma função matemática é usada para combinar os valores dos fatores e a probabilidade de *input*, a fim de produzir os índices esperados de aplicação da regra em cada célula. Diversas funções foram usadas na literatura; a que é favorecida atualmente é a "função logística"[7] (ver Rousseau & Sankoff (1978) para os detalhes matemáticos).

No texto que segue discuto cinco questões que surgem no curso desse tipo de análise. Três delas decorrem da necessidade analítica de comparação controlada: ortogonalidade dos grupos de fatores, assimetria da distribuição dos dados e sobreposição da probabilidade de *input* a algum outro fator. A quarta questão diz respeito a um problema na estimação estatística dos efeitos restritivos: decidir o que é significativo. E a última é um problema universal da estatística multivariada: independência *versus* interação.

2. Ortogonalidade dos grupos de fatores. Para fazer o tipo de comparações controladas antes descritas, os grupos de fatores devem ser "ortogonais", ou quase "ortogonais". Isto é, eles devem coocorrer livremente, e não ser sub- ou supercategorias uns dos outros. No exemplo, um falante em qualquer SEC pode usar qualquer estilo de fala; todas as combinações possíveis de classe social e estilo ocorrem nos dados. Estilo não é uma subcategoria de classe social, e um dado estilo não é exclusivamente usado por apenas um ou dois dos grupos sociais. Essa é a verdadeira ortogonalidade, que permite o maior

[5] N.T. SEC é a abreviatura utilizada para classe socioeconômica (em inglês, *socioeconomic class*).

[6] N.T. Preferi manter a palavra inglesa *input* em lugar da tradução *entrada*, devido ao seu uso generalizado no contexto da análise com o Varbrul.

[7] N.T. Na literatura estatística, encontra-se a denominação "modelo de regressão logística" neste caso.

âmbito de comparações controladas do tipo descrito acima. Mas, freqüentemente, esse ideal não é atingido. Então, temos que nos defrontar com a questão de até que ponto podemos nos desviar desse padrão e ainda obter resultados válidos.

Uma maneira de visualizar essa questão é dispor, de forma concreta, dois grupos de fatores ortogonalmente; isto é, colocá-los numa tabela bidimensional com um grupo de fatores em cada dimensão. Se fizéssemos isso com os dados nos quais a Figura 1 se baseia, todas as células da tabela seriam preenchidas, porque existem dados para todas as combinações de Estilo e SEC. Mas, num conjunto de dados com algumas células vazias, os problemas com a não-ortogonalidade poderiam surgir.

O pior caso seria o da Tabela 1, onde cada fator do grupo de fatores 1 coocorre com apenas um fator do grupo 2. Em outras palavras, todas as ocorrências de "a" são também ocorrências de "x", e todas as ocorrências de "x" são também ocorrências de "a". A mesma relação existe entre "b" e "y". Essa configuração dos fatores apresenta um problema analítico impossível de resolver. Para atribuir efeitos a "a" e "b", temos que controlar "x" e "y". Temos que poder comparar o que "a" faz quando está associado a "x" com o que "b" faz quando está associado a "x", e assim por diante, com todas as outras comparações controladas possíveis. Neste conjunto de dados, tais comparações não podem ser feitas; "a" e "x" sempre ocorrem juntos, e "b" e "y" também, de modo que seus efeitos nunca podem ser separados.

Tabela 1 - Grupos de fatores não-ortogonais: "o pior caso"

		Grupo de x	Fatores 2 y
Grupo	a	(com dados)	(sem dados)
de Fatores 1	b	(sem dados)	(com dados)

Para as finalidades do programa Varbrul, não importa se tais não-ortogonalidades ocorrem por acaso (simplesmente porque você não coletou nenhuma ocorrência nos contextos de [ay] ou [bx]), ou por impossibilidade estrutural (porque [ay] e [bx] representam impossibilidades lógicas ou combinações que não ocorrem em sua língua). Quaisquer que sejam as origens do problema, nenhum

procedimento analítico pode parcializar efeitos separados para dois grupos de fatores que apresentem esse tipo de distribuição dos dados.

Deve-se observar, contudo, que umas poucas células vazias numa tabela nem sempre criam um problema de ortogonalidade. A Tabela 1 seria analisável se, digamos, houvesse dados na célula [ay], mesmo que [bx] permacesse vazia. Ainda seria possível fazer comparações controladas de "x" e "y" (no contexto de "a") e de "a" e "b" (no contexto de "y"). Mesmo que tivéssemos menos confiança nessa análise do que em outra em que tivéssemos todas as céluas preenchidas, ainda assim poderíamos obter resultados úteis.

Essas questões podem se tornar mais concretas considerando exemplos de dados reais. Um tipo de padrão não-ortogonal é ilustrado na Tabela 2, extraída de Poplack (1980). Aqui, temos dois grupos de fatores que podem afetar o apagamento do sufixo de plural *-s* no espanhol de Porto Rico: posição da ocorrência em um SN e presença ou ausência de outras marcas de plural precedendo a ocorrência no SN. Obviamente, as combinações nessa tabela são, na maioria, impossíveis. Uma ocorrência na primeira posição (por exemplo, "las" em "las casas blancas") não pode ter quaisquer marcas de plural precedendo-a, e uma ocorrência na segunda posição só pode ter uma marca de plural a precedê-la, quer preenchida, quer apagada. Então, de fato, somente seis das 18 células da tabela apresentam dados, na medida em que essas são as únicas combinações possíveis.

Tabela 2 - Espanhol de Porto Rico: apagamento do *-s* plural; efeito da posição no SN e da presença de marca precedente (cf. Poplack, 1980)

	Posição da 1	ocorrência 2	na cadeia 3
Marca(s) precedente(s):			
'nenhuma '	0,24		
S		0,44	
0		0,52	
0S,SS			0,40
S			0,68
00			0,73

Esses "grupos de fatores", como eu os defini, são claramente não-ortogonais. Eles se aproximam da situação de "pior caso" descrita acima. Ocorrências de "primeira posição" são sempre precedidas por "nada" e ocorrências precedidas por "nada" são sempre de "primeira posição". As comparações cruzadas que gostaríamos de fazer são, na

maioria, impossíveis. Não podemos ver nunca como a "terceira posição" se comportaria quando a marca precedente fosse simplesmente "S", por exemplo. Desse modo, o procedimento analítico subjacente ao Varbrul não seria apropriado para esse tipo de codificação dos dados.

Um segundo exemplo vem do meu próprio trabalho sobre o português brasileiro. Percebi, já de início, que as variáveis que eu estava estudando apresentavam uma certa diferenciação sistemática quanto a sexo (com as mulheres favorecendo as variantes padrão mais do que os homens), mas restavam muitas diferenças idiossincráticas também. Então, num dado momento, tentei inadvertidamente analisar os dados com um grupo de fatores para "informantes" e outro para "sexo", como ilustrado na Tabela 3. A metade das células fica, obviamente, vazia, devido a combinações impossíveis. Cada informante, é evidente, tem apenas um sexo, e a análise nunca pode comparar o comportamento, digamos, do informante 6 como masculino com o mesmo informante como feminino. Conseqüentemente, os valores calculados para o grupo de fatores sexo eram sem sentido.

Tabela 3 - Grupos de fatores não-ortogonais: subcategorização

		Sexo	
		M	F
Número do Informante	1	SIM	
	2	SIM	
	3	SIM	
	4		SIM
	5		SIM
	6		SIM

Tendo entendido o problema que a não-ortogonalidade apresenta, devemos considerar como identificá-lo, quando ocorre, e como resolvê-lo. A maneira mais simples de verificá-lo é examinar cuidadosamente qualquer esquema de codificação que tenha sido estabelecido antes de empreender a análise com o Varbrul. Mas se a vigilância pessoal falha em alguma ocasião, ou se a não-ortogonalidade surge de uma possibilidade de má-distribuição dos dados não observada, então o problema se manifestará nos resultados do Varbrul como (1) resultados sem sentido e/ou (2) não-convergência. Quando qualquer uma dessas possibilidades ocorre,

deve-se interpretá-las como um aviso para verificar a não-ortogonalidade dos grupos de fatores.

Sanar esse problema é relativamente simples. Basicamente, o que queremos evitar é qualquer situação em que um fator de um grupo coocorra exclusivamente com apenas um fator de algum outro grupo. Mas devemos notar que há vários tipos diferentes de distribuição envolvidos. O pior caso, identidade completa entre dois grupos de fatores, ilustrado na Tabela 1, envolve uma relação um-a-um em ambas as direções: cada ocorrência de "a" é um "x" e cada ocorrência de "x" é um "a"; e o mesmo para "b" e "y". Quando há identidade completa entre dois grupos inteiros, ela é melhor resolvida eliminando-se, da análise, um dos grupos de fatores.

O segundo tipo de problema é a subcategorização, como na Tabela 3. Basicamente, o que acontece aqui é que cada categoria de sexo foi dividida segundo os indivíduos que as constituem, de modo que os códigos no grupo de fatores informantes são todos subcategorias de exatamente uma categoria no grupo de fatores sexo. Essa é uma relação de muitos para um: todos os dados produzidos pelo informante número um também foram codificados como tendo sido produzidos por um falante masculino, mas nem todos os dados produzidos por falantes masculinos foram produzidos pelo informante um. A solução mais simples para esse tipo de problema é, novamente, usar um só grupo de fatores constituído apenas das categorias superordenadas, neste caso, {M,F}, ou apenas das categorias subordinadas {1,2,3,4,5,6}, ou alguma combinação das duas, desde que nenhuma categoria superordenada seja usada, simultaneamente, com algumas de suas subordinadas. (Assim, um grupo composto pelos fatores {1,2,3,F} seria perfeitamente operável, mas não {1,2,3,4,F}, porque o informante 4 poderia, então, ser codificado seja como 4, seja como F.)

Finalmente, há combinações dos dois problemas, um com o outro, ou com fatores ortogonais normais nos mesmos grupos. O exemplo do espanhol na Tabela 2 combina os dois tipos de arranjos não-ortogonais (marcas precedentes codificadas como "nada" e posição "1" são idênticas, ao passo que as posições "2" e "3" têm, cada uma, diversas subcategorizações no grupo de fatores "marcas precedentes".

A resolução de tais combinações dependerá de cada caso. De novo, uma possibilidade é a redução a um só grupo de fatores. Foi

assim que Poplack obteve os valores apresentados na Tabela 2. A análise incluiu somente o grupo de fatores "marcas precedentes", sem um grupo de fatores separado para a "posição"; os resultados são apresentados em colunas separadas quanto às posições, simplesmente para ilustrar que há um padrão posicional nos dados (posições mais para o final favorecem o apagamento).

Mas suponha que sentíssemos que realmente havia dois efeitos diferentes (*i.e.* ortogonais) envolvidos num conjunto de dados como o de Poplack e não quiséssemos ignorar o padrão aparente na Tabela 2. Como poderíamos capturar isso dentro do quadro de comparações controladas necessárias para a análise multivariada? Fazê-lo envolve redefinição dos grupos de fatores. Nos dados de Poplack, por exemplo, poderíamos adotar um grupo de fatores "Posição" {posições 1,2,3} mais um grupo de fatores "Marcas precedentes", redefinido segundo qualquer uma das seguintes possibilidades:

- "sim": uma marca precedente ocorre em algum ponto da cadeia; "não": não há nenhuma marca precedente na cadeia. (Essa possibilidade combinaria as categorias de Poplack "S", "0S", "SS" e "S0" como "sim", e suas categorias "nada", "0" e "00" como "não".)

- "sim": uma marca de plural está presente na palavra imediatamente precedente; "não": não há marca de plural na palavra imediatamente precedente. (Essa possibilidade combina as categorias de Poplack "S", "SS" e "0S" como "sim", e "nada", "0", "00" e "S0" como "não".)

- "sim": uma marca de plural está presente na palavra imediatamente precedente; "não": não há marca de plural na palavra imediatamente precedente; "não se aplica" (codificado como um espaço em branco na maioria das versões do Varbrul): não existe palavra precedente. (Essa possibilidade é semelhante ao esquema de codificação prévio, exceto por preservar um *status* único para as ocorrências de primeira posição. Quando algum grupo de fatores é codificado com "não se aplica" para uma determinada célula, não se realizam cálculos para aquele grupo de fatores em relação àquela célula. Isso não viola a ortogonalidade.)

Em síntese, sempre se deve tentar criar esquemas de codificação que evitem relações um-a-um entre fatores de diferentes grupos. Se tal

distribuição ocorrer mesmo assim, resolva-a eliminando um grupo de fatores, redefinindo um grupo de fatores de modo que ele seja, de fato, ortogonal, ou usando a opção "não se aplica" para substituir um membro de um par de fatores problemático.

3. Assimetria na distribuição dos dados. Discuti a ortogonalidade apenas em termos categóricos. Mas sendo um variacionista, não posso me furtar a enfatizar que a ortogonalidade também pode ser vista como uma propriedade escalar ou relativa. Dissemos que é impossível analisar uma distribuição de dados como a que está na Tabela 1, onde as células [ay] e [bx] não contêm dados. Mas o que aconteceria se elas tivessem apenas uma ínfima quantidade de dados, como na Tabela 4? A análise Varbrul funcionaria aqui?

Tabela 4 - Assimetria na distribuição dos dados (Número de ocorrências)

		Grupo de Fatores 2	
		x	y
Grupo de Fatores 1	a	100	2
	b	2	100

A questão aqui é como os dados se distribuem entre as categorias nos diferentes grupos de fatores. O caso analítico ideal é o de uma distribuição equilibrada com todas as células contendo números iguais, como ilustrado na Tabela 5. O cenário com o pior caso, na Tabela 1, tem uma distribuição coincidente entre os fatores "a" e "x" e entre os fatores "b" e "y". No meio estão vários graus de distribuição assimétrica ou sobreposta, do que a Tabela 4 é um exemplo, com uma sobreposição de 98% entre "a" e "x" e entre "b" e "y". Então, dessa perspectiva, o problema da não-ortogonalidade se resolve com a questão do grau de sobreposição que podemos tolerar e ainda obter resultados úteis.

Tabela 5 - Distribuição equilibrada dos dados (Número de ocorrências)

		Grupo de Fatores 2	
		x	y
Grupo de Fatores 1	a	50	50
	b	50	50

Para ilustrar o que acontece com vários graus de sobreposição, preparei uma série de conjuntos de dados hipotéticos, cada um com dois grupos de fatores {a,b} e {x,y} e quatro células. Todos os conjuntos de dados tinham os mesmos índices de "aplicação da regra" por célula, conforme mostra a Tabela 6, mas diferiam quanto ao número de ocorrências alocadas às células [ax] [by] (diagonal 1), em oposição às células [ay] e [bx] (diagonal 2). O primeiro conjunto tinha uma distribuição equilibrada, isto é, uma proporção de 50:50 entre as duas diagonais. O segundo tinha uma proporção de 10:90, o terceiro de 5:95, o quarto de 1:99, e o último tinha uma distribuição coincidente (isto é, 0:100). Todos foram analisados com o Varbrul2M (uma versão para Macintosh do Varbrul2 que está sendo desenvolvida pelo autor). Os resultados estão resumidos na Tabela 7.

Tabela 6 - Percentagens de aplicação da regra para os conjuntos de dados de teste

		Grupo de Fatores 2	
		x	y
Grupo de Fatores 1	a	66,6%	50%
	b	50%	33,3%

Tabela 7 - Resultados dos conjuntos de dados de teste, mostrando o efeito da crescente assimetria na distribuição dos dados

Proporção assimetria (diag. 1: diag. 2)	Significância step-up	Significância step-down	Pesos dos Fatores a	b	x	y	Convergência (nº iter.)	Qhi^2 por célula
50:50	sim	sim	0,59	0,41	0,59	0,41	4	0
10:90	não	sim	0,58	0,42	0,58	0,42	15	0,004
5:95	não	sim	0,58	0,42	0,58	0,42	não	0,012
1:99	não	não	0,53	0,47	0,53	0,47	não	0,121
0:100	-	-	0,50	0,50	0,50	0,50	(1)	(0)

Com uma distribuição equilibrada, obtemos resultados analíticos robustos e significativos, com os valores dos fatores demonstrando, como era esperado, que os fatores "a" e "x" substancialmente favorecem a "regra". Ambos os grupos de fatores são significativos, a convergência é atingida prontamente, e o fator de erro "qui-quadrado por célula" é nulo, indicando que o modelo é perfeitamente adequado a esse conjunto de dados. Mas quando

começamos a ter distribuições altamente sobrepostas, todos esses resultados começam a cair. No nível de 90%, há um efeito do modelo sobre os grupos de fatores, a significância está começando a ficar pouco nítida, a convergência é atingida com muito mais dificuldade, e há um pequeno aumento no fator de erro. No nível de 95%, a convergência é totalmente perdida, e o fator de erro triplica. No nível de 99%, a significância também é totalmente perdida, os valores dos fatores são muito alterados, e o fator de erro novamente se multiplica, desta vez por 10. Finalmente, no nível de 100%, os valores dos fatores são sem sentido e não-significativos, e não há razão de prosseguir a análise para além da primeira iteração.

Que implicações têm esses resultados para o lingüista que esteja trabalhando com dados reais? Os detalhes vão variar em cada caso, mas podemos adotar certas "regras práticas". Uma sobreposição de 90% é tolerável, ainda que se deva reconhecer que algumas distorções dos resultados provavelmente estejam ocorrendo; além disso, a análise já está demorando muito mais para ser feita. 95% de sobreposição é, provavelmente, o limite absoluto de capacidade de análise razoável. Um pesquisador seria muito imprudente ao confiar em (ou publicar) índices baseados em uma distribuição de dados com mais de 95% de sobreposição entre dois fatores de diferentes grupos.

4. Sobreposição de dados com p_0. Um tipo de distribuição com sobreposição que freqüentemente passa despercebido, mas que pode distorcer da mesma forma os resultados analíticos, é a sobreposição entre um fator e a "probabilidade de *input*" — p_0. Isso ocorrerá sempre que houver um grupo de fatores com um fator "i" que abarca 95% ou mais dos dados, acrescidos de um ou mais fatores {"j", "k", "l"...} no grupo, os quais, juntos, compreendem menos de 5% de todas as ocorrências. Uma distribuição coincidente surge aqui, porque esse fator "i" é encontrado em 95% de todas as ocorrências, enquanto p_0, por definição, é o ambiente de 100% de todas as ocorrências. Então, quase todo o tempo, quando o algoritmo está tentando parcializar efeitos, ele não pode decidir se atribui um efeito a "i" ou a p_0.

A primeira vez que me defrontei com esse problema foi em relação a uma análise de alguns dados do português brasileiro. Eu estava estudando restrições funcionais sobre a ocorrência do *-s* indicativo de concordância de número, sob a hipótese de que os falantes omitiriam as marcas de plural mais livremente se a

informação sobre a pluralidade fosse recuperável de algum outro ponto da sentença ou do discurso. Alguns dos resultados da análise aparecem na Tabela 8.

Tabela 8 - Português brasileiro: concordância no SN; efeito do local de informação adicional de plural (cf. Guy, 1981)

	% Plural Marcado	N	Prob.
Inform. adic. plural precede SN	61	152	0,24
Sem inform. adic. plural	61	8588	0,63
Inform. adic. plural segue SN	84	1046	0,65

Considerando os valores probabilísticos dos fatores, vemos que a marcação de plural parece fortemente desfavorecida pela presença de informação adicional de plural no discurso precedente ao SN que contém a ocorrência, mas quando não há tal informação, ou quando tal informação segue o SN, a marcação de plural de uma ocorrência é favorecida. Os pesos relativos para essas duas categorias favorecedoras são muito próximos (0,63 e 0,65); de fato, um teste de significância revela que eles não são significativamente diferentes. Isso sugere que há um simples efeito binário aqui: quando a pluralidade de um SN foi claramente estabelecida no discurso antes de ele ser pronunciado, ela não terá que ser morfologicamente indicada, com freqüência, dentro do próprio SN; do contrário, as marcas morfológicas serão mais importantes.

Contudo, se implementamos tal análise, surge um grande problema de distribuição. O exame da coluna N revela que a combinação dos dois últimos fatores do grupo produz um superfator, que caracteriza 98,5% dos dados do *corpus*, em contraposição a um microfator, que compreende apenas 1,5% dos dados. O superfator é, assim, quase coincidente com a probabilidade de *input*, e todos os efeitos de distribuições sobrepostas que já foram descritos acima começam a acontecer com esta análise: valores instáveis para os fatores deste grupo e para a probabilidade de *input*, que mudaram radicalmente com pequenas revisões no esquema de codificação, impossibilidade de atingir convergência e significância, etc. A moral da história é que as mesmas regras práticas que estabelecemos para sobreposições entre dois fatores de grupos diferentes se aplicam a sobreposições entre qualquer fator e a probabilidade de *input*: não exceda os 95%.

5. Teste de significância dentro dos grupos de fatores. No trabalho recente com regras variáveis tem-se prestado bastante atenção à significância de grupos inteiros de fatores. Isso se deve, sem dúvida, à ampla disponibilidade dos programas Varbrul2 e Varbrul3, os quais, automaticamente, calculam as significâncias dos grupos de fatores através do procedimento *step-up/step-down*[8]. Mas, em minha experiência, isso seguidamente leva os usuários a negligenciarem refinamentos analíticos dentro dos grupos.

Freqüentemente, começamos uma análise com uma lista exaustiva de contextos detalhadamente discriminados, que incluem todas as distinções que podemos conceber como possivelmente relevantes. Mas muitas ou a maioria dessas distinções acabam por não ter efeito real sobre a variável que está sendo investigada. Nosso objetivo, como em qualquer empreendimento científico, deveria ser atingir uma análise que fosse EFICIENTE, GERAL e PODEROSA. Isso quer dizer que temos que identificar os fatores não significativos, que em nada contribuem para a análise, e eliminá-los. Esse é o dogma básico do método científico conhecido como "Navalha de Occam": sempre minimize o número de princípios explanatórios.

Na prática, fazer isso numa análise de regra variável envolve um duplo processo:

1. identificar as combinações de fatores que são LINGÜISTICAMENTE mais gerais e

2. verificar se são QUANTITATIVAMENTE similares, em termos de seus efeitos sobre a variação em estudo.

Assim, caso se estivesse examinando alguma variável sintática para a qual uma distinção inicial entre sujeitos, objetos diretos e objetos indiretos tivesse sido feita, poderia ser razoável tentar uma reanálise em que os dois tipos de objetos fossem reunidos, desde que os valores probabilísticos obtidos para ambos estivessem muito próximos. Seria criado um conjunto de dados recodificados, no qual esse grupo de fatores se tornaria simplesmente sujeito *versus*

[8] N.T. Mantive as expressões *step-up/step-down* do original por seu largo emprego entre os usuários dos programas Varbrul. Na literatura estatística, esses termos correspondem a métodos automáticos de seleção de modelos, denominados *stepwise*, *forward* e *backward*.

objeto, ao qual se aplicaria o teste padrão para significância. Mas não seria razoável a adoção de um procedimento como o que descrevo a seguir. Digamos que se tivesse notado que os objetos indiretos e os sujeitos tivessem valores probabilísticos similares e, então, se decidisse reuni-los numa só categoria, oposta a objetos diretos. Na falta de qualquer generalização lingüisticamente significativa sobre propriedades comuns que unissem sujeitos e objetos indiretos, esse procedimento geraria lixo, de acordo com a antiga lei da ciência da computação que diz: *garbage in, garbage out*[9].

Novamente um exemplo com dados reais servirá para iluminar a questão. A Tabela 9 apresenta alguns resultados de Guy (1981) mostrando o efeito da posição do sujeito sobre a regra variável de concordância sujeito-verbo no português brasileiro. Inicialmente, cinco categorias foram distinguidas, incluindo dois casos separados em que não havia sujeito na superfície. Na categoria 2, a referência plural do sujeito ausente é recuperável pelo complemento, em sentenças como *São cinco irmãos*. Na categoria 3, a pluralidade da referência é recuperável a partir de algum outro ponto do discurso, geralmente por continuidade de referência a um determinado tópico de discurso. (Casos em que a pluralidade da referência não pode ser estabelecida de modo algum são, necessariamente, omitidos do *corpus*, devido a um paradoxo do observador. Ver Guy (1981:243) para maiores detalhes.)

Acontece que essas duas categorias têm valores muito próximos: 0,61 e 0,55. Elas também têm muita coisa em comum do ponto de vista lingüístico: constituem todos e os únicos casos coletáveis com sujeito zero na superfície. Combiná-las em um único fator só eliminará a distinção secundária quanto a como a pluralidade da referência é estabelecida. Numa reanálise, portanto, eu reuni esses dois fatores em um e constatei, por meio do teste de *log-likelihood*[10], que a distinção assim eliminada não era significativa (p>0,70). Podemos concluir que a distinção original era irrelevante para essa regra, e sua eliminação produz um modelo mais geral e eficiente da variabilidade.

[9]N.T. Na área da computação, usa-se a expressão em inglês, daí mantê-la no original. Seu significado corresponde, aproximadamente, a "lixo que entra, lixo que sai", ou seja, se o dado colocado na entrada (no *input*) não contiver informação relevante sobre o fenômeno, o resultado (o *output*) também não será relevante.

[10] N.T. Mantive a expressão *log-likelihood* do original, por ser a forma mais freqüentemente usada no contexto; sua tradução é logaritmo da função de verossimilhança.

Tabela 9 - Português brasileiro: concordância sujeito-verbo; efeito da posição do sujeito (cf. Guy, 1981)

	% Marca de Plural	N	Probabilidades	
1. Sujeito superficial (S.S.) precede imediatamente o verbo	74	1	0,66	0,67
2. Não há S.S., mas a pluralidade é recuperável pelo complemento	68		0,61	} a }
3. Não há S.S., mas a pluralidade é recuperável pelo discurso	63	7	0,55	} b 0,51 }
4. S.S. precede o verbo e está distante	56	597	0,43	
5. S.S. segue o verbo	27	199	0,27	0,32

[a] diferença não significativa (p>0,7). [b] diferença não significativa (p>0,5)

Também podemos notar, na análise original, que a categoria 4 está relativamente próxima, quanto ao peso, das categorias 2 e 3; ao menos, próxima o suficiente de 3 para justificar o uso do teste de significância. A generalização lingüística que pode ser feita sobre as três categorias é que, enquanto nas categorias 1 e 5 há um sujeito superficial ADJACENTE ao verbo, em 2, 3 e 4 não há. Assim, mais uma reanálise foi feita, reunindo a categoria original 4 com as categorias 2&3 já combinadas. Novamente essa distinção se mostrou não significativa (p>0,50).

Feitas essas combinações, chegamos à análise mais geral e eficiente possível, que é apresentada na última coluna da Tabela 9. Um sujeito precedendo imediatamente o verbo favorece a marca de concordância de plural, um sujeito posposto desfavorece, e um sujeito apagado ou distante tem um efeito neutro.

Procedendo dessa maneira, atingimos resultados inteligíveis e gerais. Mas suponha que tivéssemos adotado a abordagem errada, agindo com base apenas na similaridade quantitativa. Os dois valores probabilísticos que mais se aproximavam, na primeira análise, eram, na verdade, os fatores 1 e 2, que diferem em apenas 0,05. Reuni-los pode ou não produzir um resultado estatisticamente significativo, mas o que a categoria combinada resultante SIGNIFICARIA? Que generalização lingüística nos permitiria reunir um sujeito imediatamente pré-verbal com um tipo de sujeito apagado, sem incluir outros sujeitos apagados ou outros sujeitos pré-verbais? Para usar um termo da fonologia, as categorias 1 e 2 não formam uma CLASSE

NATURAL, independentemente de sua similaridade numérica calculada.

Esse exemplo também ilustra o fato de que o trabalho quantitativo não é um substituto, mas apenas um acessório para a análise lingüística. O programa Varbrul apenas realiza manipulações matemáticas sobre um conjunto de dados. Ele não nos diz o que os números significam, muito menos faz lingüística por nós. Se estamos perguntando: "Qual seria a melhor generalização lingüística?", a resposta vem de nossa teoria lingüística, não de um programa estatístico.

6. Independência e interação. A análise do Varbrul difere de alguns outros tipos de análises multivariadas pelo fato de pressupor que os vários grupos de fatores têm efeitos independentes. Mas os usuários deveriam compreender claramente que ELES não devem pressupor essa independência. Ao contrário, devem ter consciência dos problemas criados pela interação, saber como detectá-la quando ocorre e saber o que fazer a respeito.

Discussões sobre como detectar e lidar com a interação podem ser encontradas em Cedergren (1973) e em outros textos da literatura. Para minimizar a repetição, vou apenas revisar brevemente os pontos que são enfatizados em outros lugares e me concentrar, aqui, em alguns outros aspectos do problema. A maneira básica de detectar a interação com o Varbrul, como ilustra Cedergren, é procurar altos valores de qui-quadrado naquelas células em que os fatores em interação coocorrem. Todas as versões do programa vão gerar uma lista de valores de qui-quadrado por célula no resultado, de modo que basta examinar essa lista em busca de valores altos (ignorando os que são artificialmente inflacionados por células de tamanho pequeno: N de 1 ou 2). Se existem vários desses valores, envolvendo coocorrência repetida de um determinado par de fatores, a interação provavelmente é a causa. O qui-quadrado é uma medida de bom ajuste entre o modelo, que pressupõe a independência, e os dados, que podem ter interação. Quando o ajuste é ruim, em geral, isso se deve à interação.

Outra técnica útil para examinar a independência de dimensões analíticas é comparar os resultados de uma análise contendo múltiplos grupos de fatores com os de uma análise contendo um só grupo de fatores. Por exemplo, se estivéssemos examinando

algum processo fonológico que fosse condicionado por consoantes adjacentes, poderíamos criar um grande grupo de fatores contendo todos os segmentos consonantais da língua, ou poderíamos ter três grupos de fatores diferentes, que classificassem cada segmento segundo Ponto, Modo e Vozeamento. De fato, a análise com três grupos testaria os vários traços distintivos separadamente, quanto a seu efeito sobre a regra, enquanto a análise com um grupo consideraria cada segmento como um todo não analisado, sem qualquer afirmação sobre classes naturais de segmentos.

Um exemplo similar envolvendo fatores sociais seria o de uma análise que tivesse um grupo de fatores para sexo e outro para classe (digamos MC versus WC)[11], em comparação com uma análise que tivesse um só grupo com quatro fatores, um para cada subgrupo da população: {homens de WC, homens de MC, mulheres de WC, mulheres de MC}.

O teste de independência neste procedimento envolve comparar os valores probabilísticos obtidos pelas duas análises. Se os índices da análise com um só grupo podem ser mapeados numa tabela, de forma sistemática, alinhados de acordo com o previsto pela análise com múltiplos grupos de fatores, não há interação. Mas, se algum subgrupo de segmentos ou subgrupo social não se alinha como esperado com base na análise de "traços" (com múltiplos grupos), então há uma possibilidade de interação. Em nossa análise de classe-por-sexo, a interação surgiria se o seguinte caso ocorresse. Suponha que a análise com dois grupos apresentasse probabilidades mais altas para homens do que para mulheres, e para falantes da classe trabalhadora do que para falantes da classe média. Então, esperaríamos que, na análise com um só grupo, o peso para homens da classe trabalhadora fosse maior do que o valor para homens da classe média, entre outras coisas. Se isso não fosse verdade, se, digamos, esses dois índices fossem iguais ou invertidos, então seria provável que classe e sexo não fossem independentes.

Um exemplo ajudará a esclarecer essas questões. Na Tabela 10 estão listadas as probabilidades obtidas em Guy (1981) para o efeito de uma consoante seguinte sobre o apagamento de *-s* final. A análise por traços mostra que as consoantes vozeadas favorecem o apagamento mais do que as não-vozeadas, me que há um efeito do

[11] N.T. MC - Middle Class/ Classe Média; WC - Working Class/Classe Trabalhadora

ponto de articulação envolvendo um pico de apagamento para dentais e um valor mínimo para velares.

Se os resultados da análise por segmentos refletisse perfeitamente os resultados da análise por traços, então cada segmento vozeado deveria ter um valor mais alto do que sua contrapartida não-vozeada no mesmo ponto de articulação. Também os valores em cada linha da tabela com a análise por segmentos deveriam ser assim ordenados: dentais > labiodentais > labiais > velares. Essas relações se mantêm para todos os valores da tabela exceto um, o segmento /g/. Esse valor é MAIS BAIXO do que o de sua contrapartida não-vozeada /k/, quando deveria ser mais alto. Com base no padrão geral, seria de esperar que /g/ tivesse um valor entre 0,35 e 0,40, em lugar de 0,12. Isso significa que deve haver, com efeito, alguma interação entre os traços [+vozeado] e [velar]. A conseqüência estatística é que a análise por traços se mostra significativamente PIOR em relação ao *log likelihood* do que a análise por segmentos (p>0,05). (É a análise por traços que é piorada pela interação, pois envolve mais restrições. Ela tenta reunir classes naturais de segmentos, ao passo que a análise por segmentos permite que o peso de cada segmento flutue livremente, sem ser restringido pela atuação de outros segmentos foneticamente similares).

Tabela 10 - Português brasileiro: apagamento de *-s* final; efeito da consoante seguinte (Probabilidades de apagamento, cf. Guy, 1981)

Análise por segmentos				Análise por Traços			
b	v	d	g	Vozeamento:			
0,47	0,63	0,72	**0,12**	segmentos vozeados:			0,58
				segmentos não-vozeados:			0,42
p	f	t	k	Ponto:			
0,38	0,45	0,53	0,27	labial	labiodental	dental	velar
				0,50	0,54	0,64	0,32

Então, tendo identificado uma interação potencial aqui, o que deveríamos fazer a respeito? Há duas respostas para essa pergunta, dependendo do objetivo que temos em mente. Analiticamente, há um recurso aceito para tratar a interação dentro do quadro do Varbrul, por meio de um termo de interação ou "traço de exceção". Isso é feito pelo acréscimo de um grupo extra de fatores à análise por traços, digamos {x,y}, em que "x" é codificado em cada célula contendo os

traços [+vozeado] e [velar] (*i.e.* cada célula contendo o caso excepcional de /g/), e "y" é codificado em todas as outras células. A análise resultante atribuirá o valor excepcionalmente baixo de /g/ ao fator "x".

A razão para se fazer isso é que os valores obtidos SEM usar um traço de exceção serão distorcidos pela interação. Se /g/ é excepcionalmente baixo, então os valores calculados para [+vozeado] e [velar] também serão demasiado baixos, a fim de atingir um ajuste estatístico melhor entre o modelo e os dados. Quando a natureza excepcional de /g/ é acomodada pela inclusão de um termo de interação no modelo, uma estimativa mais acurada dos valores dos outros fatores pode ser obtida.

A segunda resposta à pergunta "o que fazer" é teórica — trata-se de investigar e comprovar que os resultados estão errados ou são suspeitos, e que realmente não há interação. Faz-se isso quando se está confiante, com base na teoria ou em outros aspectos independentes, de que não deveria haver qualquer interação entre os termos em questão e se suspeita dos resultados. No presente caso, eu não pude achar qualquer explicação razoável para que /g/ se comportasse de forma peculiar no condicionamento do apagamento de *-s*, o que me levou a duvidar desse resultado. Se fosse um caso de interação entre alguns parâmetros sociais como, digamos, sexo e ocupação, eu não teria a mínima desconfiança. Mas o resultado em questão era teoricamente improvável, por isso, busquei explicações alternativas.

Uma possível resposta foi encontrada na quantidade de dados neste caso. Os números de ocorrências (N) correspondentes à tabela 10 estão na tabela 11. Aí vemos que o segmento desviante /g/ tem, de longe, o menor número de ocorrências sustentando os resultados analíticos. Então, neste caso, apesar da significância estatística no nível de 0,05, eu decidi não me preocupar com isso. É preciso lembrar que, neste nível de significância, ainda se espera que um caso em vinte seja produzido por flutuação puramente aleatória dos dados. Em outras palavras, há uma chance em vinte de que o valor de /g/ anormalmente baixo tenha surgido porque aconteceu, de forma bastante aleatória, que o conjunto de *-s* finais seguidos de palavras começadas por /g/ que encontramos incluíam, apenas por acaso, um reduzidíssimo número de apagamentos. Neste exemplo, minhas expectativas teóricas eram fortes o suficiente para que eu interpretasse o resultado como sendo aquele único caso em vinte. Obviamente, essa

visão pode ser testada por replicação: coletar mais casos, digamos, mais 75 ou 100 ocorrências de *-s* final antes de /g/ e ver se a mesma taxa reduzida de apagamentos prevalece. Se isso ocorrer, o fenômeno é real e requer explicação; caso contrário, minhas desconfianças estarão confirmadas. Deixo essa tarefa a algum leitor interessado.

Tabela 11 - Português brasileiro: apagamento de *-s* final; número de ocorrências por consoante seguinte (cf. Guy, 1981)

b	v	d	g
122	278	736	75
p	f	t	k
497	314	498	961

7. Conclusões. O Varbrul é uma ferramenta poderosa e extremamente útil para a análise da variação lingüística. Como acontece com qualquer ferramenta, sua utilidade é acentuada por uma compreensão de suas operações e de suas limitações. Mas não se deve nunca perder de vista o fato de que, na análise final, o Varbrul é apenas um recurso (embora sofisticado) para a manipulação dos dados. Não discerne padrões, não faz generalizações, nem explica achados. Isso tudo é com você.

Referências Bibliográficas

CEDERGREN, Henrietta J. *The interplay of social and linguistic fators in Panama*. Dissertation, Cornell University, 1973.

_____& SANKOFF, David. Variable rules: Performance as a statistical reflection of competence. In: *Language*, 1974, 50: 233-55.

GUY, Gregory R. Use and applications of the Cedergren/Sankoff variable rule program. In: FASOLD, Ralph W. & SHUY, Roger W. (eds.) *Analyzing variation in language*. Washington: Georgetown University Press, 1975. p.59-69.

_____. *Linguistic variation in Brazilian Portuguese: Aspects of the phonology, syntax, and language history*. Dissertation, University of Pennsylvania, 1981.

LABOV, William *The social stratification of English in New York City*. Arlington, VA: Center for Applied Linguistics, 1966.

_____. Contraction, deletion, and inherent variability of the English copula. In: *Language*, 1969, 45:715-762.

POPLACK, Shana. The notion of plural in Puerto Rican Spanish: Competing constraints on /s/ deletion. In: LABOV, William. (ed.) *Locating language in time and space*. New York: Academic Press, 1980. p. 55-68.

ROUSSEAU, Pascale. *A versatile program for the analysis of sociolinguistic data*. Université de Montréal, Centre de Recherche de Mathematiques Apliquées, 1983. (CRMA-1163).

_____& SANKOFF, David. Advances in variable rule methodology. In: SANKOFF, David. (ed.) *Linguistic variation: models and methods*. New York: Academic Press, 1978. p. 57-69.

SANKOFF, David. *VARBRUL version 2*. [Unpublished manuscript]. 1975.

_____ & LABOV, William. On the uses of variable rules. In: *Language in Society*, 1979, 8:189-222.

WEINREICH, Uriel, LABOV, William & HERZOG, Marvin. Empirical foundations for a theory of language change. In: LEHMANN, Winfred & MALKIEL, Yakov. *Directions for historical linguistics*. Austin: University of Texas Press, 1968. p. 97-195.