Codificação estatística das categorias fonéticas: vestígio da dinâmica da fala na fonotaxe lexical[1]

Statistical encoding of phonetic categories:
speech dynamics traces on lexical phonotactics

Eleonora C. Albano[2]
LAFAPE,[3] IEL,UNICAMP

Abstract

This paper presents a new, intriguing finding and discusses some of its theoretical implications: phonetic categorization, including major class membership, is entirely predictable from phonotactic biases in three Brazilian Portuguese word databases. The predictors are log frequencies of 'VC, 'CV and V'CV sequences consisting of the 7 stressed vowels combined with the 19 onset consonants, plus the 5 Southeastern pre-stressed vowels. Correct vowel categorization arises through discriminant analysis of 'VC and 'CV data. Correct consonant categorization arises through discriminant analysis of V'CV data. Results are consistent across databases and, thus, strongly suggest that statistical biases in the lexicon can be so stable as to code phonetic categories. The findings and their corollaries bear on the issue of the relationship of lexical phonotactics to speech dynamics.

Keywords

Phone statistics, Dynamics, Phonotactics, Lexicon, Brazilian Portuguese

Resumo

Este artigo apresenta um novo e intrigante achado e discute algumas de suas conseqüências teóricas: a categorização fonética, inclusive no tocante às grandes classes, é inteiramente previsível a partir de vieses fonotáticos em três bases de dados lexicais do português brasileiro. Os preditores são as freqüências logarítmicas de seqüências ‘VC, ‘CV e V’CV em que as 7 vogais tônicas se combinam às 19 consoantes mediais de ataque, mais as 5 vogais pretônicas do sudeste. A categorização das vogais emerge via análise discriminante dos dados de ‘CV e ‘VC. A categorização das consoantes emerge pela mesma via dos dados de V’CV. O fato de os resultados serem consistentes entre as bases sugere que vieses estatísticos lexicais podem ser estáveis o bastante para codificar categorias fonéticas. Os achados e seus corolários afetam a questão da relação entre a fonotaxe lexical e a dinâmica da fala.

Palavras-chave

Estatística fônica, Dinâmica, Fonotaxe, Léxico, Português brasileiro

1. Introdução

Há muito se sabe que a sintaxe dos fones ou fonotaxe não é uma combinatória arbitrária e exibe vieses distribucionais foneticamente motivados. Um exemplo é a escala de sonoridade, que leva a maioria das línguas a preferir a ordem obstruinte/soante ou o seu reverso em encontros consonantais de ataque e coda silábica, respectivamente (CLEMENTS, 1990). Há também encontros consonantais de certa complexidade motora que são proibidos em famílias inteiras de línguas sem que se possam dizer impronunciáveis: por exemplo, */dl/ nas línguas românicas. Proibições mais específicas e até conflitantes refletem estratégias de produção distintas entre línguas diferentes: por exemplo, a ocorrência exclusiva de vogais abertas sob nasalidade em francês e a tendência, diametralmente oposta, do português, brasileiro ou europeu, a excluí-las em contextos de nasalização.

Mais recentemente, as relações fonotáticas deixaram de ser vistas apenas como categóricas. Seguindo uma sugestão de Greenberg (1950), Pierrehumbert (1994, 2003) propôs a existência de restrições fonotáticas probabilísticas. Desde então, vem-se estudando, em diversas línguas, as funções lexicais e gramaticais dessa fonotaxe ampliada – que tem se revelado não só gradiente, mas também, em certos casos, não seqüencial (V. FRISCH, 1996, para o árabe; ALBANO, 2003, para o português).

Tanto quanto aqui se tenha notícia, Maddieson (1993) foi o primeiro a sugerir que vieses fonotáticos podem constituir pistas para a natureza dos primitivos fônicos lexicais. Com base em dados fonotáticos interlingüísticos, argumentou que os vieses CV são o reflexo de uma economia de gestos do corpo da língua. A mesma idéia foi explorada foneticamente por Traunmüller (1999).

Este trabalho retoma a preocupação com a fonotaxe, para mostrar, com dados do léxico do português brasileiro (doravante PB), que a economia da combinatória fônica é tal que é capaz de codificar, via estatística fônica, todas as categorias pertinentes para a classificação das vogais e consoantes da língua,

inclusive as chamadas grandes classes (soante, líquida, etc.).

Os dados provêm de três bases de dados lexicais do PB contemporâneo, foneticamente codificadas e de tamanhos que variam entre aproximadamente 25.000 e 225.000 palavras. Nelas foram calculadas e analisadas as freqüências logarítmicas de todas as seqüências 'VC, 'CV e V'CV.

A metodologia de análise envolve as seguintes técnicas estatísticas multivariadas: o escalonamento multidimensional, para a exploração inicial dos dados; a análise discriminante, para a análise propriamente dita; e a análise de grupamentos (*cluster analysis*), para a interpretação dos resultados.

Os achados são inequívocos, robustos e se podem dizer esperáveis no tocante às categorias compartilhadas por vogais e consoantes, isto é, as que envolvem os mesmos articuladores, a saber: o ponto de articulação nas consoantes; a posição da língua e o arredondamento dos lábios nas vogais. São, por outro lado, surpreendentes no tocante às demais categorias, a saber: as de abertura e posição da raiz da língua nas vogais; e as de modo de articulação e grandes classes nas consoantes. Ressalte-se que nenhuma das últimas envolve propriedades vocálicas e consonantais obviamente comuns.

Depois de uma discussão metodológica que se impõe em virtude da falta de precedentes na literatura, os resultados serão apresentados por classe, isto é, vogal e consoante, numa ordem que vai dos mais esperados aos mais surpreendentes. Considerações sobre as variáveis independentes selecionadas pela análise discriminante para resolver as classificações aparentemente mais opacas abrirão o caminho para a defesa de uma organização lexical consistente com restrições provenientes dos mecanismos de produção e percepção de fala.

Ao final, será possível argumentar que, por não serem logicamente necessários, tais reflexos da dinâmica da fala só se tornam inteligíveis se interpretados como marcas deixadas no léxico pelos primitivos fônicos constitutivos das suas entradas. Será defendida a visão de que eles são gestuais e de que a incorporação fonológica de informação motora abstrata não exclui, mas até, ao contrário, requer a incorporação paralela de informação auditiva, como uma espécie de elo de controle sensório-motor. O grau de abstração dessas entradas lexicais foneticamente motivadas é, por sua vez, compatível com a tarefa de portar informação gramatical.

2. Metodologia

2.1. As bases de dados: CETENFolha, NURC e MiniAurélio

A maior amostra analisada provém de um *corpus* público denominado CETENFolha[4] (doravante CETEN). Trata-se de um conjunto de textos de 24 milhões de palavras do jornal *A Folha de São Paulo,* coletado e tratado pelo *Núcleo Interinstitucional de Lingüística Computacional* (NILC), da USP, e disponibilizado pela *Linguateca*, sítio interinstitucional de Internet sediado em Portugal e dedicado ao estudo do processamento computacional da língua portuguesa.

Até recentemente não havia material de língua oral que pudesse ser comparado ao CETEN. A situação mudou no final de 2006, quando se terminou de preparar para esse fim uma parte do acervo do Projeto NURC,[5] já utilizada em Albano *et al.* (1995), que contém 45.579 palavras. Era também desejável comparar listas de palavras extraídas de textos a um dicionário. Para tanto, utilizou-se o MiniAurélio, de 27.074 palavras, primeira base de dados lexicais informatizada construída para o PB (SILVA *et al*, 1994).

Mesmo na Lingüística de *Corpus*, área que vem crescendo muito graças aos progressos da informática, a questão do tamanho é, ainda hoje, bastante controversa. Por causa do mais conhecido *corpus* do inglês americano (FRANCIS; KUCERA, 1982), havia, até bem pouco tempo, uma prática consensual de considerar um milhão de palavras suficiente. Hoje se sabe que o bom tamanho é relativo: pode ser maior ou menor, a depender do foco da pesquisa. A razão é que as distribuições de freqüência de ocorrência de palavras têm um caráter zipfiano, isto é, crescem exponencialmente (ZIPF, 1949) com o aumento do vocabulário. Assim, *corpora* relativamente pequenos, embora não forneçam amostras representativas de palavras de baixa freqüência, podem ser satisfatórios para as de alta ou média, conforme os objetivos da análise.

É compreensível, portanto, que, na Estatística Fônica, área em que se insere este trabalho, a questão do tamanho seja ainda mais obscura, tendo em vista que a literatura sobre a freqüência de ocorrência de fones é escassa mesmo nas línguas mais estudadas. Curiosamente, os estudos mais importantes vêm se produzindo não na Fonética ou na Fonologia, mas na Psicolingüística e na Psicologia Experimental (VITEVICH; LUCE, 1999), nas quais as probabilidades fonotáticas se têm afirmado como fatores a serem controlados em experimentos.

Assim, decidiu-se reportar aqui primariamente os resultados do CETEN, que, além de ser o maior *corpus*, foi o único a passar na maioria dos testes de

compatibilidade com os pressupostos da principal técnica estatística empregada, a análise discriminante. Como se verá a seguir, essa decisão respalda-se também (1) nas altas correlações[6] entre os valores das variáveis independentes em todos os *corpora*; e (2) nos espaços fônicos semelhantes que eles projetam a partir de um tratamento estatístico multivariado. Ressalte-se, porém, que o MiniAurélio e o NURC foram indispensáveis para apoiar considerações sobre tamanho e tipo de base de dados, a saber: língua oral ou escrita; ou, ainda, *corpus* ou dicionário.

2.2. Tratamento dos *corpora*

O *software* Ortofon, de nossa autoria (ALBANO; MOREIRA, 1996), fez a conversão fônica automática da transcrição ortográfica dos três *corpora*. A intenção inicial era converter a saída desse programa em caracteres do *Alfabeto Fonético Internacional*, moeda comum a toda a comunidade da Lingüística. Entretanto, o programa estatístico utilizado para as análises e para a confecção dos gráficos não reconhece esses caracteres. Optou-se, então, pela conversão da saída do Ortofon em SAMPA, o alfabeto alternativo proposto pela Associação Fonética Internacional para uso com o computador.[7]

O primeiro *corpus* a ser tratado foi o CETEN, cujos resultados já foram divulgados em parte anteriormente (ALBANO, 2005, 2006, 2007). As unidades iniciais de análise foram os pares ‘CV e ‘VC, compreendendo as 7 vogais tônicas e as 19 consoantes de ataque silábico medial do PB,[8] o que produziu 133 pontos de dados para cada caso. A insatisfação com o desempenho dessas duplas na análise das consoantes levou à adição da tripla V’CV, que combina as 5 vogais pretônicas[9] aos mesmos fones, produzindo 665 pontos de dados. Toda a contagem foi feita com o *software* Freq, elaborado por Fráguas (2005) a partir da saída do Ortofon na versão revista por L. C. F. Oliveira (2003) para fins de uso com outros programas. O referido programa, disponível na página de internet do LAFAPE,[10] retorna as freqüências, computadas sobre os tipos e as ocorrências das palavras do *corpus*, para fones ou seqüências fônicas, especificados na entrada conforme as convenções da versão do Ortofon utilizada. Um total de aproximadamente 15 milhões de palavras, distribuídas por 223.193 tipos, prestou-se à análise das unidades ‘CV e ‘VC. Dessas, apenas cerca de 5 milhões de ocorrências, distribuídas entre 126.449 tipos, atenderam às condições para o cômputo de V’CV.

No NURC, a contagem foi feita de maneira semi-automática, sobre a saída regular do Ortofon, através de recursos do programa *Excel*. No Mini-

Aurélio, ela fez uso do *software* elaborado originalmente para tanto, denominado Listas (SILVA *et al., op. cit.*).

2.3. Análise estatística

Todas as freqüências obtidas foram convertidas em logaritmos e submetidas a análises estatísticas com o *software* Statistica 6.0.

Três técnicas exploratórias multivariadas – o escalonamento multidimensional (*multidimensional scaling*), a análise discriminante (*discriminant analysis*) e a análise por grupamentos (*cluster analysis*) – foram utilizadas para revelar os grupamentos possíveis entre os fones de acordo com os seus vieses de co-ocorrência. No CETEN e no NURC, em que se fez também um cômputo de ocorrências de palavras, o subconjunto do *corpus* que melhor se prestou às análises foi, não obstante, o de tipos, o que facilitou a comparação com o MiniAurélio. Em todos os *corpora*, 'CV e 'VC prestaram-se coerente e economicamente à análise das vogais, enquanto V'CV apresentou os melhores resultados para as consoantes.

As técnicas estatísticas multivariadas servem para explorar e extrair estruturas de dados sempre que a comparação entre as variáveis é dificultada pela existência de múltiplas relações entre elas. É, portanto, desejável que os seus resultados sejam consistentes com análises conceituais realizadas prévia ou subseqüentemente.

O escalonamento multidimensional é uma ferramenta quantitativa baseada em matrizes de distância[11] e se presta a extrair dimensões ortogonais de um espaço multivariado, isto é, constituído por duas ou mais variáveis, e aí alocar os objetos estudados. É uma alternativa à análise fatorial, com a vantagem de ser mais flexível, por não ser sensível à normalidade.

A análise discriminante é uma espécie de análise de variância às avessas, isto é, avalia os efeitos de variáveis independentes quantitativas sobre variáveis dependentes qualitativas. Assim, dado um conjunto de medidas, permite não só determinar se elas discriminam entre dois ou mais grupos ou categorias naturais, mas também selecionar aquelas que melhor o fazem.

A análise por grupamentos também se baseia em métricas de distância, para daí extrair diagramas em árvore, sem pressupor normalidade. Ela se presta a agrupar objetos em categorias de tal forma que a associação entre quaisquer dois deles seja máxima se pertencerem ao mesmo grupo e mínima no caso contrário. O formalismo das árvores permite, ainda, hierarquizar as categorias de acordo com o mesmo critério.

3. Discussão metodológica

Nos estudos de Estatística Lingüística é uma prática de consenso converter as freqüências de ocorrência brutas das unidades inventariadas em logaritmos, a fim de comprimir as diferenças entre os dados da faixa mais alta e aproximar a distribuição, geralmente exponencial, a alguma outra forma mais tratável – de preferência, a normal. Essa prática traz um problema quando há, entre as referidas unidades, casos cuja freqüência é zero ou um. Como se sabe, o primeiro número não tem logaritmo e o segundo tem por logaritmo zero. Nos *corpora* aqui tratados, há zeros em 'VC quando C é nasal e, também, 0 ou 1 em muitas combinações de V'CV. A Tabela 1 resume essa situação:

TABELA 1
Ocorrências das freqüências brutas 0 e 1 em 'VC e V'CV nos três *corpora*

	CETEN		NURC		MiniAurélio	
	Freq. 0	Freq. 1	Freq. 0	Freq. 1	Freq. 0	Freq. 1
'VC	3	0	3	0	3	0
V'CV	2	12	54	39	64	58

É evidente que a informação contida na Tabela 1 sobre a rejeição do PB a certas combinações de vogais e consoantes não pode ser simplesmente nivelada ou, pior ainda, ignorada, como faria automaticamente a maioria dos programas de análise estatística.[12] Assim, para tornar esses casos tratáveis pelas mesmas técnicas que os demais, foi preciso adicionar uma constante a todas as células. Após alguns testes do efeito das candidatas sobre a forma da distribuição, o valor escolhido foi 0,1.

Outra consideração metodológica indispensável diz respeito aos pressupostos subjacentes às técnicas estatísticas utilizadas. Enquanto o escalonamento multidimensional e a análise por grupamentos, como já foi dito, não pressupõem normalidade, a análise discriminante o faz, embora tolere desvios moderados. Dos três *corpora*, o único que se aproxima de uma distribuição normal das variáveis e suas combinações – a chamada normalidade multivariada – é o CETEN.

Isso, por si só, não impediria o único *corpus* de língua oral, o NURC, de ser a referência central do estudo se os demais pressupostos da análise discriminante nele estivessem contemplados. Mas, conforme esperado, só o CETEN, devido à grande amostra, apresenta desvios de fato negligenciáveis da homoscedasticidade ou homogeneidade de variância – a condição de todas as variáveis terem

variâncias estatisticamente iguais, exigida pela análise discriminante, devido à sua lógica subjacente, baseada na regressão linear.[13]

Uma razão ainda mais forte para impedir que a exposição seja centrada nos resultados do NURC é que, no subconjunto de tipos, as médias e variâncias da unidade V'CV são correlacionadas para os 35 pares de vogais. A correlação entre esses parâmetros afeta as regressões lineares envolvidas nos procedimentos em questão e é, portanto, fatal para a sua confiabilidade.

Assim, optou-se por tomar o CETEN como fio da exposição, uma vez que as suas correlações com os demais *corpora* são altíssimas para todos os parâmetros estudados. Devido aos desvios da normalidade dos dados das bases menores, foram calculados também os coeficientes não-paramétricos de correlação de Spearman e Kendall, com resultados congruentes. Assim, a Tabela 2 limita-se a ilustrar a semelhança entre os *corpora* com o coeficiente de correlação de Pearson.

TABELA 2
Coeficientes de correlação de Pearson entre os dados do CETEN e dos demais *corpora*

Nível de Sig: p<0,05	Seqüência	MiniAurélio	NURC
CETEN	V_V em V'CV	0,96	0,98
CETEN	C em V'CV	0,96	0,99

Infere-se da Tabela 2 que as diferenças entre o PB escrito e oral são mínimas no que toca aos vieses de coocorrência entre consoantes e pares de vogais tônicas e pretônicas. Cabe lembrar que a seqüência V'CV é o lugar de maior contraste lexical possível, na medida em que abriga o inventário máximo de vogais, o da tônica, o inventário máximo de consoantes, o da posição medial, e o inventário máximo de vogais átonas no dialeto estudado, o da posição pretônica. O fato de os três *corpora* apresentarem vieses de coocorrência altamente coerentes entre si indica que o léxico do PB, ao invés de maximizar o referido contraste (o que minimizaria os vieses de coocorrência), regula-o através de restrições cuja origem pode, em princípio, ser ou puramente estocástica, ou fônico-gramatical, como aqui se argumenta.[14]

4. Resultados

Antes de passar aos resultados propriamente ditos, é importante ressaltar a representatividade do CETEN para fins de estudo dos vieses lexicais de co-ocorrência, tendo em vista tratar-se de um *corpus* de língua escrita.

4.1. Semelhanças entre os *corpora*: Escalonamento multidimensional

Para confirmar e realçar as semelhanças entre o CETEN e o NURC, usou-se o escalonamento multidimensional, computado a partir de matrizes de distâncias euclidianas[15] entre as freqüências logarítmicas. As Figuras 1 e 2 exibem espaços consonantais que refletem diretamente os vieses de ocorrência das 19 consoantes com os 35 pares V'_V nos dois *corpora*.

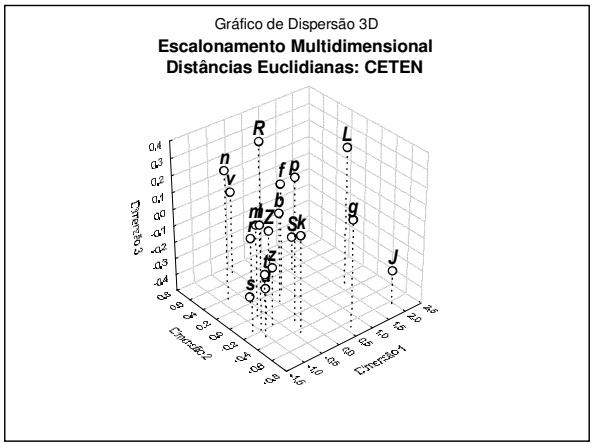

FIGURA 1 - Gráfico 3D das distâncias euclidianas entre as freqüências logarítmicas de C em V'CV no CETEN

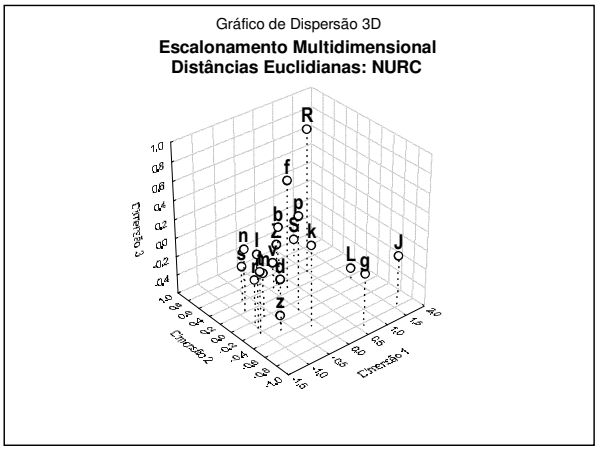

FIGURA 2 - Gráfico 3D das distâncias euclidianas entre as freqüências logarítmicas de C em V'CV no NURC

É importante ressaltar que, apesar de algumas superposições gráficas, há, em ambos os casos, evidentes e comparáveis agrupamentos fonéticos, os quais parecem privilegiar, embora não cabalmente, o ponto de articulação. A separação quase perfeita entre consoantes anteriores e não-anteriores é digna de nota.

O uso do CETEN para garantir a confiabilidade das técnicas estatísticas empregadas está, portanto, suficientemente respaldado. É de esperar que um aumento do tamanho da amostra do NURC mantenha a coerência com os padrões de coocorrência aqui observados, aproximando-o ainda mais do CETEN.

4.2. Classificação fonética das freqüências de coocorrência

A estabilidade dos gradientes de coocorrência constituídos pelas freqüências logarítmicas de 'CV, 'VC e V'CV nos três *corpora* corrobora a hipótese de que o léxico do PB abriga princípios combinatórios de natureza fônica e/ou gramatical. Obviamente, explorar a via fônica é mais simples, até porque ela não exclui, mas, ao contrário, às vezes até recobre e sinaliza a gramatical. Assim, é possível indagar, por exemplo, se tais gradientes são foneticamente interpretáveis. A técnica estatística adequada para tanto é a análise discriminante (doravante, AD).

A AD agrupa dados quantitativos por meio de regressões lineares e compara os grupos obtidos com uma classificação *a priori*, a fim de calcular, entre outros parâmetros, a sua taxa de acerto e medidas de distância entre os grupos, as quais são submetidas a testes de significância. Para isso, é necessário eliminar redundâncias excessivas, pela exclusão de variáveis altamente correlacionadas, através de um parâmetro denominado "tolerância.[16] Quando é possível fazê-lo manualmente de maneira unívoca, pode-se usar a técnica "passo a passo" (*stepwise*) para alcançar uma solução econômica, com a melhor taxa de acerto para o menor número de variáveis. Nos casos em que é preciso escolher entre variáveis muito numerosas e correlacionadas, a técnica indicada é a denominada "melhor subconjunto" (*best subset*). Ela permite especificar os tamanhos mínimo e máximo do subconjunto, o critério de ordenação das soluções e um parâmetro que ajuda a evitar empecilhos ao cálculo causados pelo excesso de redundância.[17]

Nos dados do CETEN, o tamanho do subconjunto ótimo variou entre três e seis variáveis. Cabe lembrar que o número inicial dessas variáveis é 19, no caso das vogais, classificadas através da coocorrência com as consoantes mediais de ataque; e 35, no caso das consoantes, classificadas através da coocorrência com os pares de vogais pretônicas e tônicas.

É importante também ressaltar que a coocorrência com as sete vogais tônicas presentes em ‘CV e ‘VC não foi produtiva para a classificação das consoantes: as taxas de acerto dos melhores subconjuntos, tanto para os pontos como para os modos de articulação, variaram, com ambas as unidades, entre 73% e 89%, o que é pouco em comparação com os 100% obtidos em todas as análises apresentadas a seguir.

4.2.1. As classes vocálicas

‘CV e ‘VC produzem resultados semelhantes para os descritores vocálicos relacionados ao lugar da constrição, mas não para os relacionados à abertura do trato vocal. Esses são corretamente diferenciados pelos melhores modelos baseados em ‘VC, mas não em ‘CV. Nesse caso, a taxa máxima de acerto foi de 71% para a abertura propriamente dita, com duas variáveis apenas, já que os modelos de três variáveis não passavam nos critérios de tolerância. Analogamente, para a posição da raiz língua, a taxa máxima de acerto em ‘CV foi de 84%, também com um modelo de duas variáveis, pela mesma razão.

Reportam-se abaixo os melhores modelos encontrados com taxa de acerto de 100%, tendo em conta: o número de variáveis; a tolerância obtida para cada uma; e os níveis de significância alcançados, no todo e nas comparações aos pares.

4.2.1.1. Posição da língua

Um resultado padrão da AD é a matriz de classificação, conforme se vê adiante. Ela exibe a taxa de acerto, total e por categoria, bem como as probabilidades *a posteriori* das categorias,[18] cuja soma deve aproximar-se de 1. Foram classificadas *a priori* como anteriores as vogais /i, e, E/; e, como posteriores, as vogais /a, O, o, u/. Essa classificação foi 100% reproduzida na análise.

TABELA 3
Matriz de classificação da AD para a posição da língua das vogais em 'CV

AD: Matriz de Classificação			
Posição da Língua	Linhas: classificações *a priori* Colunas: classificações previstas		
	% Correto	Anterior p=0,42857	Posterior p=0,57143
Anterior	100	3	0
Posterior	100	0	4
Total	100	3	4

Outra informação importante dada pela AD é o lambda de Wilks, número que varia entre 0 (classificação perfeita) e 1 (ausência de classificação) e que desempenha um papel equivalente ao do F na análise de variância. Assim como esse, associa-se sempre a um nível de significância, que é a probabilidade de se cometer um erro estatístico do Tipo I, isto é, rejeitar uma hipótese nula verdadeira. A hipótese nula aqui seria a ausência de diferenciação entre vogais anteriores e posteriores por meio dos seus vieses de coocorrência em 'CV. Na Tabela 4 e congêneres, é também reportado o lambda de Wilks parcial, isto é, referente à contribuição específica de cada variável, ao lado do F correspondente.

TABELA 4
Modelo de três variáveis extraído pela AD para a posição da língua das vogais em 'CV

AD: Variáveis no Modelo: 3						
Consoantes Preditoras (N=7)	Agrupamento: Posição da Língua (2 grupos) Lambda de Wilks: 0,01145 aprox. F (3,3) = 86,307 p< 0,0021					
	Lambda de Wilks	Lambda Parcial	F a remover (1,3)	p	Tolerância	1-Tolerância (R^2)
R	0,046396	0,246873	9,152	0,056524	0,046894	0,953106
Z	0,059383	0,192879	12,5538	0,038278	0,077773	0,922227
g	0,439644	0,026052	112,1523	0,001799	0,020093	0,979907

Cabe notar que os R^2 da Tabela 4 são altíssimos, o que quer dizer que as três consoantes incluídas são muito correlacionadas entre si. É interessante também notar que os pontos de articulação não-anteriores estão mais representados que os anteriores: há uma palatal e uma velar, ao lado do rótico forte, que foi classificado *a priori* – e também pela AD – como dental, conforme se verá em 4.3.1.

Nas conclusões, será retomada a possível interpretação fonética dessa seleção de variáveis, feita pela AD por critérios puramente computacionais.[19]

4.2.1.2. Arredondamento

No tocante ao arredondamento, foram classificadas *a priori* como não-arredondadas as vogais /i, e, E, a/; e, como arredondadas, as vogais /O, o, u/.

TABELA 5
Matriz de classificação da AD para o arredondamento dos lábios das vogais em 'CV

AD: Matriz de Classificação			
	Linhas: classificações *a priori* Colunas: classificações previstas		
Arredondamento dos Lábios	% Correto	Não-arredondada P=0,57143	Arredondada P=0,42857
Não-arredondada	100	4	0
Arredondada	100	0	3
Total	100	4	3

Observe-se que, nesse caso, as três consoantes selecionadas são anteriores, sendo duas dentais e uma labial. Todas são, além disso, obstruintes vozeadas.

TABELA 6
Modelo de três variáveis extraído pela AD para o arredondamento dos lábios das vogais em 'CV

AD: Variáveis no Modelo: 3						
Consoantes Preditoras (N=7)	Agrupamento: Arredondamento dos lábios (2 grupos) Lambda de Wilks: 0,01928 aprox. (3,3) = 50,864 p< 0,0045					
	Lambda de Wilks	Lambda Parcial	F a remover (1,3)	p	Tolerância	1-Tolerância (R^2)
v	0,748838	0,025748	113,5136	0,001767	0,027081	0,972919
d	0,181792	0,106061	25,2855	0,015155	0,048326	0,951674
z	0,195095	0,098830	27,3553	0,013599	0,068313	0,931687

4.2.1.3. Abertura

Conforme já mencionado, a abertura ou altura da língua não alcançou a taxa máxima de acerto em 'CV. Em 'VC, entretanto, o seu desempenho foi muito bom. Assumiram-se três aberturas ou alturas vocálicas. As vogais /i, u/ foram classificadas como fechadas ou altas; as vogais /e, o/, como médias; e as vogais /E, a, O/, como abertas ou baixas.

TABELA 7
Matriz de classificação da AD para a abertura ou altura da língua das vogais em 'VC

AD: Matriz de Classificação				
Abertura ou Altura da Língua	Linhas: classificações *a priori* Colunas: classificações previstas			
	% Correto	Fechada p=0,28571	Média p=0,28571	Aberta p=0,42857
Fechada	100	2	0	0
Média	100	0	2	0
Aberta	100	0	0	3
Total	100	2	2	3

TABELA 8
Modelo de 3 variáveis extraído pela AD para a abertura ou altura da língua das vogais em 'VC

AD: Variáveis no Modelo: 3						
Consoantes Preditoras (N=7)	Agrupamento: Abertura ou Altura da Língua (3 grupos) Lambda de Wilks: 0,00034 aprox. F (6,4)=35,742 p< 0,0020					
	Lambda de Wilks	Lambda Parcial	F a remover (1,3)	p	Tolerância	1-Tolerância (R^2)
b	0,039789	0,008426	117,6745	0,008426	0,072555	0,927446
r	0,027873	0,012029	82,1326	0,012029	0,013249	0,986751
m	0,015162	0,022113	44,2217	0,022113	0,027980	0,972020

É importante notar o predomínio reiterado de consoantes anteriores: aqui, curiosamente, as labiais. Não obstante, um rótico aparece de novo – desta vez o brando, inequivocamente dental (detalhes nas seções 5 e 6). Além disso, todas as consoantes são vozeadas, sendo duas soantes e uma obstruinte.

Cabe, finalmente, nesse caso, reportar também os níveis de significância das comparações aos pares, já que se trata de uma distinção tríplice.

TABELA 9
Níveis de significância das comparações aos pares dos graus de abertura extraídos pela AD

p aos pares	Fechada	Média	Aberta
Fechada		0,010580	0,018266
Média	0,010580		0,035648
Aberta	0,018266	0,035648	

4.2.1.4. Posição da raiz da língua

Além de estar longe de 100% de acerto em 'CV, este foi o único descritor vocálico a exigir um modelo de quatro variáveis em 'VC. Nenhum dos modelos possíveis de três variáveis atingiu o nível de significância mínimo. A posição "raiz avançada" englobou, *a priori*, as vogais /i, e, o, u/; a posição "raiz retraída" englobou as vogais /E, a, O/. Essa classificação é parcialmente redundante com a de abertura, mas tem a vantagem de ressaltar o que as vogais médias têm em comum com as fechadas ou altas.

TABELA 10
Matriz de classificação da AD para a posição da raiz da língua das vogais em 'VC

AD: Matriz de Classificação			
Posição da Raiz da Língua	Linhas: classificações *a priori* Colunas: classificações previstas		
	% Correto	Avançada p=0,57143	Retraída p=0,42857
Avançada	100	4	0
Retraída	100	0	3
Total	100	4	3

Cabe notar a ausência de palatais e velares na Tabela 11. O modelo é, de novo, composto exclusivamente de consoantes anteriores – nesse caso, três dentais e uma labial. Há duas soantes e duas obstruintes, ambas desvozeadas. Chama atenção também a repetição do rótico brando e da nasal labial.

TABELA 11
Modelo de quatro variáveis extraído pela AD para a posição da raiz da língua das vogais em 'VC

AD: Variáveis no Modelo: 4						
Consoantes Preditoras (N=7)	Agrupamento: Posição da Raiz da Língua (2 grupos) Lambda de Wilks: 0,01017 aprox. F (4,2)=48,684 p< 0,0202					
	Lambda de Wilks	Lambda Parcial	F a remover (1,2)	p	Tolerância	1-Tolerância (R^2)
r	0,39706	0,025603	76,11615	0,012884	0,021629	0,978371
t	0,045854	0,221699	7,02122	0,117787	0,0375	0,9625
s	0,014676	0,692687	0,88731	0,445642	0,089838	0,910162
m	0,448911	0,022646	86,317	0,011388	0,02109	0,97891

4.3. As classes consonantais de ataque silábico

No caso das consoantes, conforme já mencionado, os modelos baseados em 'CV e 'VC atingiram taxas de acerto relativamente baixas mesmo com sete variáveis, isto é, todas as vogais. Em contraste, os melhores modelos baseados em V'CV não só atingiram a taxa de acerto de 100%, mas também mantiveram o número de variáveis entre quatro e seis. Isso indica que pares de vogais são mais informativos do que vogais isoladas sobre as consoantes com que coocorrem. As pistas dadas pelos resultados a seguir para possíveis explicações para esse fato serão retomadas na conclusão.

4.3.1. Ponto de articulação

Foram classificadas *a priori* como labiais /p, b, f, v, m/; como dentais, /t, d, n, s, z, l, r, R/; como palatais, /S, Z, J, L/; como velares, /k, g/. Essa classificação foi 100% reproduzida pela AD. Cabe relatar, a propósito, que também foi testada uma tentativa de classificar o rótico forte como velar quanto ao ponto e fricativa quanto ao modo de articulação (V. 4.3.2) e que o mau desempenho da AD apontou claramente para a superioridade da classificação conservadora, isto é, dental e líquida.

TABELA 12
Matriz de classificação da AD para o ponto de articulação das consoantes em V'CV

AD: Matriz de Classificação Linhas: classificações *a priori* Colunas: classificações previstas					
Ponto de articulação	% Correto	Labial p=0,26316	Dental p=0,42105	Velar p=0,10526	Palatal p=0,21053
Labial	100	5	0	0	0
Dental	100	0	8	0	0
Velar	100	0	0	2	0
Palatal	100	0	0	0	4
Total	100	5	8	2	4

TABELA 13
Modelo de cinco variáveis extraído pela AD para o ponto de articulação das consoantes em V'CV

AD: Variáveis no Modelo: 5						
Pares V'_V	Agrupamento: Ponto de Articulação (4 grupos) Lambda de Wilks: 0,01059 aprox. F (15,30)=8,6035 p< 0,0000					
Preditores (N=19)	Lambda de Wilks	Lambda Parcial	F a remover (3,11)	p	Tolerância	1-Tolerância (R^2)
o_e	0,055499	0,190736	15,55716	0,000285	0,156198	0,843802
u_e	0,037070	0,285556	9,17380	0,002493	0,154401	0,845599
o_E	0,088721	0,119313	27,06475	0,000022	0,145012	0,854988
e_u	0,032317	0,327553	7,52746	0,005175	0,214340	0,785660
o_u	0,047994	0,220560	12,95768	0,000624	0,141325	0,858675

Salta aos olhos que todos os pares têm ao menos uma vogal média. São também internamente distintos quanto à posição da língua e ao arredondamento dos lábios, à exceção do último.

Cabe, nesse caso, reportar também os níveis de significância das comparações aos pares, visto tratar-se de uma distinção quádrupla.

TABELA 14
Níveis de significância das comparações aos pares dos pontos de articulação extraídos pela AD

p aos pares	Labial	Dental	Velar	Palatal
labial		0,009147	0,000008	0,017445
dental	0,009147		0,000055	0,000059
velar	0,000008	0,000055		0,000002
palatal	0,017445	0,000059	0,000002	

4.3.2. Modo de articulação

É importante ressaltar que também as classificações clássicas de modo de articulação, seja quanto ao tipo de constrição, seja quanto à presença ou à natureza do vozeamento, são integralmente replicadas pela AD.

4.3.2.1. Tipo de constrição

Foram classificadas *a priori* como oclusivas /p. b, t, d, k, g/; como fricativas, /f, v, s, z, S, Z/; como nasais, /m, n, J/; como líquidas, /l, r, R, L/. Essa classificação foi 100% reproduzida pela AD.

TABELA 15
Matriz de classificação da AD para o modo de articulação das consoantes em V'CV

AD: Matriz de Classificação					
Linhas: classificações *a priori* Colunas: classificações previstas					
Ponto de articulação	% Correto	Oclusiva p=,31579	Fricativa p=,31579	Nasal p=,15789	Líquida p=,21053
Oclusiva	100	6	0	0	0
Fricativa	100	0	6	0	0
Nasal	100	0	0	3	0
Líquida	100	0	0	0	4
Total	100	6	6	3	4

TABELA 16
Modelo de cinco variáveis extraído pela AD para o modo de articulação das consoantes em V'CV

AD: Variáveis no Modelo: 5						
Pares V'_V	Agrupamento: Ponto de Articulação (4 grupos) Lambda de Wilks: 0,01382 aprox. F (15,30)=7,6224 p< 0,0000					
Preditores (N=19)	Lambda de Wilks	Lambda Parcial	F a remover (3,11)	p	Tolerância	1-Tolerância (R^2)
o_i	0,075675	0,182634	16,40997	0,000225	0,080804	0,919196
a_e	0,075447	0,183185	16,34957	0,000229	0,075069	0,924931
u_a	0,072799	0,189849	15,64693	0,000278	0,184822	0,815178
a_O	0,052763	0,261942	10,33135	0,001571	0,23634	0,76366
a_u	0,040852	0,338313	7,17143	0,006143	0,147109	0,852891

Chama a atenção a presença da vogal /a/ em todos os pares, à exceção do primeiro. Cabe notar, além disso, que todos diferem em arredondamento, à exceção do segundo.

Reportam-se a seguir os níveis de significância das comparações aos pares – visto tratar-se, de novo, de uma distinção quádrupla.

TABELA 17
Níveis de significância das comparações aos pares dos modos de articulação extraídos pela AD

p aos pares	Oclusiva	Fricativa	Nasal	Líquida
Oclusiva		0,008719	0,003900	0,004161
Fricativa	0,008719		0,000374	0,007566
Nasal	0,003900	0,000374		0,000210
Líquida	0,004161	0,007566	0,000210	

4.3.2.2. Natureza ou presença do vozeamento

A AD também tem sucesso com as distinções binárias de modo, seja entre soantes e obstruintes, seja entre vozeadas (ou sonoras) e desvozeadas (ou surdas).

4.3.2.2.1. Obstruência

Foram classificadas como soantes /m, n, l, r, R, J, L/ e, como obstruintes, /p, b, f, v, t, d, s, z, S, Z, k, g/. A distinção é em parte redundante com a de vozeamento, pois a manutenção da voz durante a constrição oral, embora não requeira esforço nas soantes, exige manobras para reduzir a pressão supraglótica nas obstruintes vozeadas ou sonoras. A AD, mais uma vez, alcançou 100% de acerto.

TABELA 18
Matriz de classificação da AD para a obstruência das consoantes em V'CV

AD: Matriz de Classificação			
Obstruência	Linhas: classificações a priori Colunas: classificações previstas		
	% Correto	Obstruinte p=0,63158	Soante p=0,36842
Obstruinte	100	12	0
Soante	100	0	7
Total	100	12	7

Essa foi a distinção consonantal que resultou no menor modelo, a saber, de quatro variáveis.

TABELA 19
Modelo de quatro variáveis extraído pela AD para a distinção de obstruência das consoantes em V'CV

AD: Variáveis no Modelo: 4						
Pares V'_V	Agrupamento: Obstruência (2 grupos) Lambda de Wilks: 0,22919 aprox. F (4,14)=11,771 p< 0,0002					
Preditores (N=19)	Lambda de Wilks	Lambda Parcial	F a remover (1,14)	p	Tolerância	1-Tolerância (R^2)
a_e	0,377968	0,606364	9,08844	0,009277	0,325094	0,674906
a_E	0,380769	0,601903	9,25954	0,008772	0,094402	0,905598
e_u	0,50684	0,452187	16,96065	0,001044	0,131231	0,868769
a_u	0,754275	0,30385	32,07539	0,000059	0,038078	0,961922

Note-se a repetição de /a_e/ e /a_u/. Observe-se também que os pares são formados por quatro vogais, /e, E, a, u/, em combinações que portam distinções de abertura e posição da língua. Exceto pelo terceiro, eles carregam

também distinções quanto à posição da raiz da língua. Cabe, finalmente, notar a presença do /a/ em três dos quatro pares.

4.3.2.2.2. Vozeamento

A distinção de vozeamento, conforme já mencionado, só é pertinente para as obstruintes. Assim, foram classificadas como desvozeadas ou surdas /p, f, t, s, S, k/ e, como vozeadas ou sonoras, /b, v, m, d, z, n, l, r, R, Z, L, J, g/.

TABELA 20
Matriz de classificação da AD para a distinção de vozeamento das consoantes em V'CV

AD: Matriz de Classificação			
Vozeamento	Linhas: classificações *a priori* Colunas: classificações previstas		
	% Correto	Desvozeada p=0,31579	Vozeada p=0,68421
Desvozeada	100	6	0
Vozeada	100	0	13
Total	100	6	13

Também aqui a AD obteve 100% de acerto. Curiosamente, porém, esse é o modelo consonantal que exigiu o maior número de variáveis: seis, pois o melhor modelo de cinco variáveis alcançou apenas 94% de acerto.

TABELA 21
Modelo de seis variáveis extraído pela AD para o vozeamento das consoantes em V'CV

AD: Variáveis no Modelo: 6						
Pares V'_V	Agrupamento: Vozeamento (2 grupos) Lambda de Wilks: 0,15268 aprox. F (6,12)=11,099 p< 0,0003					
Preditores (N=19)	Lambda de Wilks	Lambda Parcial	F a remover (1,12)	p	Tolerância	1-Tolerância (R^2)
u_e	0,371135	0,411384	17,16986	0,001362	0,215105	0,784895
a_a	0,524663	0,291004	29,23661	0,000158	0,084916	0,915084
o_a	0,866926	0,176115	56,13731	0,000007	0,017007	0,982993
i_O	0,616618	0,247607	36,46395	0,000059	0,079065	0,920935
a_o	0,307857	0,49594	12,19649	0,004444	0,165502	0,834499
o_o	0,714057	0,213819	44,12235	0,000024	0,032866	0,967134

Predominam aqui as vogais posteriores, sendo /a/ e /o/ repetidas mais de uma vez. Chamam a atenção, a propósito, os dois pares de vogais idênticas, /a_a/ e /o_o/. Todos os demais têm vogais distintas quanto ao arredondamento. Metade dos pares exibe harmonia quanto à posição da raiz da língua.

5. Vieses fonéticos nas seleções de variáveis da AD

A exposição anterior deixou claro que todos os subconjuntos selecionados pela AD, seja de consoantes a classificar as vogais, seja de pares de vogais a classificar as consoantes, têm bastante consistência fonética. Para focalizar apenas os dois exemplos mais intuitivos, lembremos que as três consoantes que melhor classificaram a posição da língua nas vogais, a saber, /R, Z, g/, implicam o uso desse articulador – ponta, para /R/ (V. 6), e corpo, isto é, frente e dorso, para /Z/ e /g/, respectivamente; de maneira análoga, os pares de vogais que melhor classificaram o ponto de articulação das consoantes são, à exceção de um, distintos justamente quanto à posição da língua.

Esses fatos parecem banais quando se considera que, *a priori*,[20] 73,7% das consoantes do PB fazem uso da língua como articulador, assim como 48,5% dos pares de pretônicas e tônicas consistem de vogais distintas quanto à posição da língua. Mas é justamente essa enorme redundância da fonotaxe da posição tônica que permite a extração de informação fonética por meio de meras freqüências de coocorrência. Nos outros dois *corpora* estudados, isto é, no NURC e no MiniAurélio, as variáveis selecionadas pela AD, embora não sejam exatamente as mesmas, têm características fonéticas semelhantes.

A lógica subjacente às seleções é transparente: segmentos foneticamente semelhantes têm freqüências de coocorrência ainda mais correlacionadas entre si do que os demais. Assim, na última seção, a variação de R^2, na última coluna das Tabelas de 4 a 19, está entre 0,674 a 0,983. Isso indica que todas as variáveis selecionadas pela AD são bastante correlacionadas entre si. Ora, essas correlações são ainda mais altas entre algumas variáveis que, por pertencerem a mais de uma categoria fonética em comum, acabaram excluídas – devido à redundância excessiva. Há uma métrica simples que permite traduzir esse tipo de afinidade em termos de distâncias: trata-se de 1-R (onde R é o coeficiente de correlação de Pearson). Quanto mais correlacionadas as variáveis, menor é esse valor e a distância entre elas. Examinemos, nas Figuras 3 e 4, dois exemplos, um relativo a vogais e outro relativo a consoantes, das hierarquias que a análise por grupamentos é capaz de construir por meio dessas distâncias.

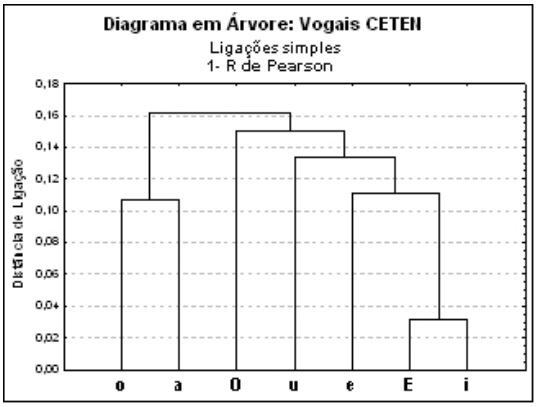

FIGURA 3 - Diagrama em árvore das distâncias 1- R de Pearson entre as freqüências logarítmicas de V em 'CV no CETEN

Observe-se que, pelo menos nos ramos inferiores, a separação entre vogais anteriores e posteriores é bem-sucedida. Analogamente, abaixo, algumas distinções de ponto e de modo de articulação entre as consoantes foram corretamente captadas através das distâncias 1-R.

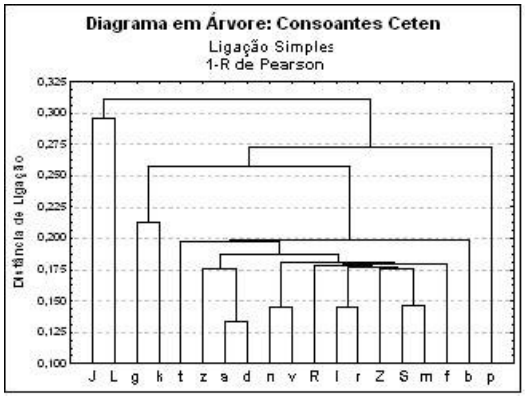

FIGURA 4 - Diagrama em árvore das distâncias 1-R de Pearson entre as freqüências logarítmicas de C em V'CV no CETEN

Destaquem-se, aqui, quanto ao ponto de articulação, a concentração de labiais, à direita, e a de dentais, ao centro. A harmonia desses conjuntos só é quebrada pelo agrupamento errôneo de /n/ e /v/, único desvio claro da motivação

fonética nesses dados. As velares estão, por sua vez, corretamente agrupadas um pouco mais à esquerda. Já as palatais cindem-se em dois grupos, relativamente distantes, mas coerentes quanto ao modo de articulação (soantes *vs*. obstruintes). Finalmente, cabe ressaltar que três das quatro líquidas estão corretamente agrupadas e de maneira coerente com os seus pontos de articulação.

Isso deve ter bastado para mostrar que os dados fonotáticos sobre os quais a AD operou nas seções anteriores contêm vieses fonéticos. Tais vieses não seriam esperados se a distribuição da informação no léxico fosse independente da influência dos mecanismos de produção e percepção da fala.

6. Conclusões

Antes de passarmos às implicações fonológicas desses dados, é oportuno traduzir o espaço consonantal da Figura 1 nos termos da métrica de distância introduzida na última seção. Os dados são os mesmos da árvore da Figura 4, agora tratados pelo escalonamento multidimensional.

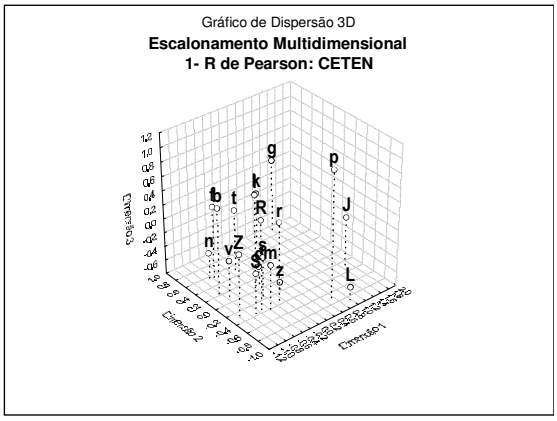

FIGURA 5 - Gráfico 3D das distâncias 1-R de Pearson entre as freqüências logarítmicas de C em V'CV no CETEN

Embora as Figuras 1 e 5 sejam visualmente muito semelhantes, esta expressa mais transparentemente o fato de as distâncias entre as 19 consoantes se basearem na sua coocorrência com as variáveis independentes – que são, nesse caso, os 35 pares de vogais. Diz também, explicitamente, que as consoantes mais próximas entre si têm freqüências logarítmicas altamente correlacionadas

no ambiente V'_V, ou seja, tendem a preferir ou rejeitar os mesmos pares de vogais. Em suma, a Figura 5 constitui uma espécie de resumo de tudo que se tentou mostrar até agora sobre a consistência fonética de tal espaço, projetado – repita-se – com base em dados exclusivamente estatísticos fonotáticos.

Reflitamos, então, sobre as implicações do tipo de redundância lexical aqui abordado. É importante frisar que se trata de uma fonotaxe puramente quantitativa, pois as 665 possibilidades de combinação de vogais e consoantes em V'CV de fato ocorrem no PB, embora algumas sejam extremamente raras.[21] Por que combinações de fones em princípio igualmente lícitas teriam vieses favoráveis ou desfavoráveis e, além disso, foneticamente motivados?

Não há como fugir à conclusão de que o léxico "olha para frente", isto é, obedece, na medida do possível, a restrições ditadas pelos mecanismos de produção e percepção da fala (DAVIDSON, 2006).

Um bom exemplo relativo à produção envolve as líquidas anteriormente destacadas. Sabe-se que no PB, assim como em outras línguas, essa classe é foneticamente muito heterogênea (SILVA, 1999). Os róticos, em particular, apresentam grande variação dialetal e socioletal quanto ao ponto e ao modo de articulação (SILVA, 2002). Não obstante, a AD foi integralmente capaz de alocá-los às classes corretas com base nas freqüências logarítmicas. Houve, até, um achado surpreendente: o rótico forte, claramente fricativo e não-anterior no dialeto do sudeste – base da transcrição ortográfico-fônica – tem uma fonotaxe não só de líquida, mas também de dental. Essa classificação conservadora, que evoca as formas diacrônicas, deve-se somente aos seus vieses de coocorrência com vogais.

A Figura 6 apresenta uma expressão simples e direta dos vieses 'V em V'LV no CETEN. Para cada par de líquida e vogal tônica, as freqüências brutas observadas foram comparadas às esperadas caso as ocorrências das vogais aí se mantivessem proporcionais às do *corpus* todo. O resultado é expresso através da razão O/E, cujos valores indicam vieses favoráveis acima de 1 e desfavoráveis abaixo de 1. Está claro que /l, r, R/ se opõem a /L/ quanto à coocorrência com a vogal /i/: essa tende a ser favorecida pelos primeiros e evitada pelo último.

Esse padrão faz perfeito sentido à luz de estudos articulatórios como os de Sproat e Fujimura (1993) e Espy-Wilson (2004), os quais revelam, nas líquidas do inglês, um gesto vocálico aproximante subjacente – semelhante a [j], [w] ou [3'] – sincronizado ao gesto consonantal característico do rótico ou da lateral. Obviamente, no caso de /L/, o gesto que produz o [j] é audível, porque não inteiramente sincronizado ao que forma a corrente de ar lateral. A rejeição por

/i/ é, portanto, nesse caso, um efeito do princípio do contorno obrigatório (doravante OCP[22]). Analogamente, nas demais líquidas, a atração por essa vogal "encoberta" é um efeito do princípio do menor esforço (doravante PME),[23] num contexto em que, aliás, não há violação audível de OCP.

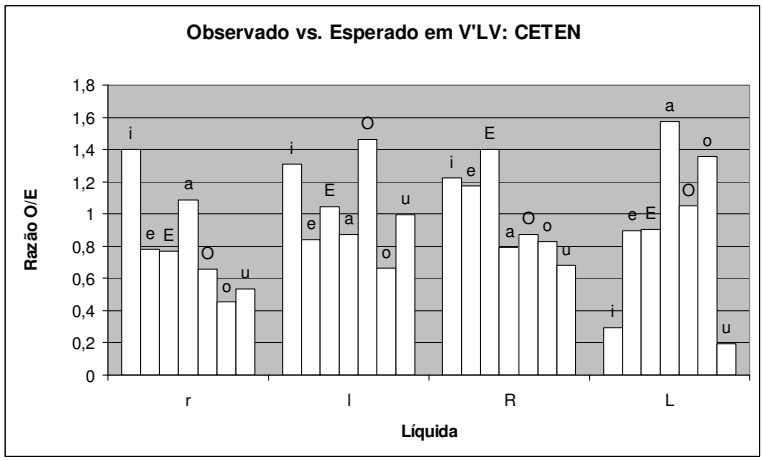

FIGURA 6 – Coocorrência das 7 vogais tônicas do PB com as 4 líquidas, expressa em termos de desvios do esperado se as respectivas ocorrências se mantivessem proporcionais ao total em V'CV no CETEN.

A Figura 6 torna ainda mais compreensíveis as posições de /L/ nas Figuras 4 e 5. A maior ou menor coocorrência antes de /i/ parece ter sido a principal responsável pela sua separação e pelo concomitante agrupamento de /l, r, R/. Observe-se, por outro lado, que a rejeição, nesse contexto, de /u/, assim como de uma ou mais vogais de raiz da língua retraída, pode ter tido um papel importante no sucesso da AD em identificar as líquidas como classe.

Um bom exemplo relativo à percepção diz respeito às sibilantes /s, z, S, Z/. Um funcionamento puramente abstrato de OCP quanto ao uso dos lábios como articulador cria a expectativa de que as consoantes mais desafetas aos pares de duas vogais arredondadas sejam as labiais, a saber: /p, b, f, v, m/. Ao contrário, no CETEN, os pares /o_O, u_O, o_o, u_o, o_u, u_u/ são ligeiramente preferidos por essa classe, o que sugere que o princípio em jogo, nesse caso, é PME. Curiosamente, porém, eles são fortemente rejeitados pelas sibilantes, nas quais, de maneira não óbvia, OCP atua por razões primariamente perceptuais. A Tabela 22 exibe os dados relevantes:

TABELA 22
Observado e esperado para a coocorrência de sibilantes, labiais e outras C com pares V'_V distintos quanto ao arredondamento no CETEN

V_'V	$\div^2 = 208{,}1551$, graus liberdade = 7, $p < 0{,}000000$ $\ddot{o}^2 = 0{,}040573$			
	Consoantes	Observado	Esperado	O/E
Ambas as Vogais Arredondadas	s, z	414	708,8	0,5841
	S, Z	82	151,7	0,540482
	labiais	999	838,3	1,191675
	outras	2881	2677,2	1,07613
Outros Pares	s, z	20067	19772,2	1,014909
	S, Z	4302	4232,3	1,016473
	labiais	23225	23385,7	0,993129
	outras	74479	74682,8	0,997271
	Soma	126449	126449,0	

Notem-se os baixos valores da razão O/E para /s, z. S, Z/. Uma versão articulatória de OCP não poderia explicar esses dados. Entretanto, uma versão acústico-articulatória[24] do mesmo princípio ilumina-os de maneira surpreendente. O que está em jogo aqui é o papel do gesto de protrusão labial em realçar o contraste acústico entre /s, z/ e /S, Z/. Keyser e Stevens (2006) demonstram que esse gesto é usado na produção de /S, Z/ em várias línguas por causar o abaixamento das ressonâncias do ruído fricativo e alinhá-lo com o terceiro formante da vogal seguinte, realçando o contraste com /s, z/. De fato, no PB, /S, Z/ são normalmente arredondados. É interessante observar, na Tabela 22, que tanto as sibilantes dentais como as palatais rejeitam o ambiente vocálico duplamente arredondado. Isso se deve, naturalmente, à atuação gradiente, no léxico do PB, de uma versão de OCP que busca evitar o mascaramento do referido gesto de realce. Note-se que esse efeito tende fatalmente a ocorrer sempre que um par arredondado é ladeado por /s, z, S, Z/, pois a influência de PME leva à manutenção do arredondamento durante o ruído.

Esses dois exemplos devem ter bastado para dar uma idéia de como é possível um léxico "olhar para frente". Ora, isso traz à baila a questão dos primitivos fônicos lexicais: traços abstratos ou gestos articulatórios? É evidente que toda a argumentação anterior pesa em favor dos gestos (BROWMAN; GOLDSTEIN, 1992), ainda que exija a sua especificação acústica ou auditiva.

Como não há espaço para aprofundar esse argumento agora, basta mencionar uma única vantagem de um léxico de base gestual – a qual pode ter ressonâncias no setor da comunidade de fonólogos brasileiros que se dedica aos estudos de variação e mudança lingüística (e.g., OLIVEIRA, 2003) ou aquisição da linguagem (e.g., CRISTÓFARO-SILVA e GOMES no prelo). Trata-se do caráter necessário da sílaba e da palavra como unidades fonológicas.

O gesto articulatório só faz sentido como primitivo fônico quando se reconhece a sua inelutável tendência a agregar-se pela exigência de co-produção (GOLDSTEIN; FOWLER, 2003). Unidades do tamanho da sílaba são inescapáveis, assim como é inescapável a emergência de fenômenos da ordem da acentuação quando elas se agrupam, pois, nesse processo, algumas naturalmente se enfraquecem em função do fortalecimento de outras.

Por serem oscilações abstratas que se materializam no trato vocal, os gestos articulatórios naturalmente assumem uma organização hierárquica, como, aliás, é próprio de toda conduta sensório-motora.[25]

Assim, fenômenos fônicos idiomáticos, isto é, específicos de certas palavras, são esperados, a depender de fatores como a freqüência de co-ocorrência e a afinidade dos gestos componentes. Também é esperado que algumas inovações surgidas na fala corrente e inicialmente circunscritas a certos ambientes venham a formar, às vezes, esquemas mais abstratos capazes de se generalizar e se espalhar, no léxico ou na comunidade. Tanto quanto outros tipos de modismo sensório-motor, esses processos costumam se propagar pela prática repetida e inconsciente.

A ótica gestual, desse modo, naturalmente aproxima e, portanto, trata como compatíveis e complementares, os fenômenos da difusão lexical e da mudança fonética regular.

Cabe, por fim, observar que gradientes fonotáticos como os apresentados constituem fatores de coesão lexical e podem contribuir para a maior ou menor permeabilidade de um ambiente à mudança. Por outro lado, a sua estabilidade e coerência entre diferentes bases de dados apontam para um grande potencial de conservar a informação "subjacente" em detrimento de mudanças "superficiais" – como, aliás, foi aqui demonstrado para o caso do rótico forte.

Notas

[1] Versão revista da conferência de abertura do *IX Encontro Nacional de Fonética e Fonologia*, intitulada "Modelos Fônicos Dinâmicos: para uma Contribuição Brasileira", proferida em 27 de novembro de 2006, no auditório da reitoria da UFMG. Embora o título tenha sido modificado para adequar-se ao formato de artigo, a questão da contribuição brasileira aos modelos dinâmicos continua em pauta e é abordada na conclusão.

[2] Agradeço aos organizadores do evento, César Reis e Rui Rothe-Neves, e à presidente da Sociedade Brasileira de Fonética, Mirian da Matta Machado, pelo honroso convite e demais sinais de apreço. Agradeço, ainda, a um revisor anônimo desta revista pelas valiosas correções e sugestões.

[3] Esta pesquisa contou com o generoso apoio da *FAPESP, Fundação do Amparo à Pesquisa do Estado de São Paulo* (processos no. 01/00136-2 e no 03/09564-2) e do *CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico* (processo no. 304621/2003-0).

[4] Eis a explicação encontrada em http://www.linguateca.pt: "o CETENFolha (**C**orpus de **E**xtractos de **T**extos **E**lectrónicos **N**ILC/**Folha** S. Paulo) é um *corpus* de cerca de 24 milhões de palavras em português brasileiro, criado pelo projecto *Processamento computacional do português* (projecto que deu origem à Linguateca) com base nos textos do jornal *Folha de S. Paulo* que fazem parte do *corpus* NILC/São Carlos, compilado pelo *Núcleo Interinstitucional de Lingüística Computacional* (NILC)".

[5] Transcrição ortográfica de 72 horas de gravação de 58 homens e 43 mulheres de cinco capitais brasileiras.

[6] A correlação é uma medida da independência de duas variáveis, geralmente expressa como a soma dos produtos dos seus desvios normalizados em torno da média dividida pelo número de graus de liberdade.

[7] Ver nota 8 abaixo.

[8] A saber: /p, b, m, f, v, t, d, n, s, z, l, r, R, J, s, z, J, S, Z, L, k, g/; /i, e, E, a, O, o, u/; /i, e, a, o, u/. O alfabeto fonético aqui utilizado é o SAMPA, cujas convenções para o português se encontram em http://www.phon.ucl.ac.uk/home/sampa/portug.htm. Cabe aqui apontar, de qualquer modo, as equivalências IPA dos caracteres maiúsculos, a fim de facilitar a leitura: /R/ = o rótico forte, de pronúncia variável; /J/ =/J/, /S/ = / S/, /Z/ = /Z/, /L/ = /L/, /E/ =/E/, /O/ =/O/.

[9] O programa Ortofon faz uma conversão ortográfico-fônica larga de acordo com a pronúncia do PB do sudeste.

[10] http://www.lafape.iel.unicmp.br.

[11] Por exemplo: a euclidiana, a baseada em correlações, etc.

[12] Os dois tratamentos possíveis para lacunas em computações que não as tolerem são eliminar a linha que as contém ou preenchê-las com a média das observações. Ambos os procedimentos obviamente mascarariam vieses estatísticos importantes para este estudo.

[13] A regressão linear é o ajuste de uma reta a um conjunto de pontos definidos pelos valores de ao menos duas variáveis.

[14] Note-se que aí reside, não raro, o fim do radical.

[15] Trata-se de distâncias geométricas num espaço multidimensional, calculadas pela seguinte fórmula: distance(x,y) = $\{Ó_i (x_i - y_i)^2 \}^{½}$

[16] A tolerância é a diferença entre 1 e a correlação múltipla entre a variável dada e as demais do modelo. Quanto mais se aproxima de zero, mais redundante é a variável.

[17] Esse é o problema denominado "mau condicionamento da matriz": uma tolerância abaixo de 0,01 indica uma correlação acima de 99% e impede a inversão da matriz de correlação ou covariância, que é parte do cálculo da AD.

[18] As probabilidades *a priori* são calculadas com base na atribuição das categorias pelo pesquisador; as probabilidades *a posteriori* são calculadas com base no modelo de análise.

[19] A computação do melhor subconjunto de variáveis se faz pela escolha daquele que resulta no menor lambda de Wilks juntamente com níveis de tolerância que evitem o mau condicionamento da matriz.

[20] Isto é, consideradas apenas as combinações possíveis e não o número efetivo das suas ocorrências no léxico.

[21] As duas únicas triplas ausentes do CETEN, /e'JE/ e /u'RO/ parecem ser lacunas lexicais. A segunda, porém, ocorre ao menos em "burróide". De qualquer forma, são comuns entre palavras: por exemplo, 'sem essa', 'baú rosa'.

[22] Sigla referente à consagrada nomenclatura de língua inglesa: *obligatory contour principle*.

[23] Cf. Zipf (1949) e Traunmüller (2001).

[24] Talvez fosse menos polêmico dizer "auditivo-articulatória". Mas, aqui, como em Albano (2001), prefiro chamar atenção para as relações acústico-articulatórias no trato vocal. A idéia é a de que, por se fundamentarem em princípios físicos gerais, elas seriam controladas pelo falante de maneira multimodal.

[25] Esta noção de sensório-motricidade não implica um nível inferior de conduta, desprovido de relação com a cognição abstrata. Ao contrário, a perspectiva gestual supõe que não haja uma fronteira entre o sensório-motor e o cognitivo; e que a ação motora faça parte do processo contínuo de construção da cognição.

Referências

ALBANO, E. C. Representações dinâmicas e distribuídas: indícios do português brasileiro adulto e infantil. *Letras de Hoje*, 2007.

ALBANO, E. C. Simple quantitative phonotactic indices for major phonetic classes. Pôster apresentado na *X^th Conference on Laboratory Phonology*, Paris, Sorbonne, 29 de junho a 1º. de julho de 2006. Resumo disponível em: <http://aune.lpl.univ-aix.fr/~labphon10/>.

ALBANO, E. C. Sobre o abrimento 3 de Mattoso Câmara: pistas fonotáticas para a classe das líquidas. *Estudos da Língua(gem)* v. 2, p. 45-66, 2005.

ALBANO, E. C. The interplay of phonetics and grammar in determining V-to-V phonotactics. *Proceedings of the XV^th International Congress of Phonetic Sciences*, Barcelona, 3-9 ago. 2003. p. 2393-2396.

ALBANO, E. C. *O Gesto e Suas Bordas: Esboço de Fonologia Acústico-Articulatória do Português Brasileiro*. Campinas: Mercado de Letras, 2001.

ALBANO, E. C.; MOREIRA, A. Archisegment-based letter-to-phone conversion for concatenative synthesis in Portuguese. *Proceedings of ICSLP' 96*, v. 3, 1996. p. 1708-1711.

ALBANO, E. C.; MOREIRA, A.; AQUINO, P.; SILVA, A.; KAKINOHANA, R. Segment frequency and word structure in Brazilian Portuguese. *Proceedings ICPhS' 95*, v. 3, 1995. p. 346-349.

BROWMAN, C. P.; GOLDSTEIN, L. Articulatory Phonology: an overview. *Phonetica* 49, 3-4, p. 155-180, 1992.

CLEMENTS, G. N. The role of the sonority cycle in core syllabification. In: Beckman, M.; Kingston, J. (Org.). *Papers in Laboratory Phonology I*. Cambridge: Cambridge University Press, 1990. p. 283-333.

CRISTÓFARO-SILVA, T.; GOMES, C. A. Aquisição fonológica na perspectiva multirepresentacional**.** *Letras de Hoje*, no prelo.

DAVIDSON, L. Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics 34*, p. 104-137, 2006.

ESPY-WILSON, C. Articulatory strategies, speech acoustics and variability. Comunicação ao evento *"From Sound to Sense"*, MIT, jun. 2004.

FRÁGUAS, C. C. Relatório Científico Final. Bolsa de Treinamento Técnico vinculada ao Projeto Temático *Integrando Parâmetros Contínuos e Discretos em Modelos do Conhecimento Fônico e Lexical*. FAPESP, abr. 2005.

FRANCIS, N. W.; KUCERA, H. *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston: Houghton-Mifflin, 1982.

FRISCH, S. *Similarity and Frequency in Phonology.* 1996. Tese (Doutorado) – Northwestern University, 1996. (Inédita)

GOLDSTEIN, L.; FOWLER, C. A. Articulatory phonology: a phonology for public language use. In: SCHILLER, N. O.; MEYER, A. (Org.). *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*. Berlim: Mouton de Gruyter, 2003. p. 159-207.

GREENBERG, J. The patterning of root morphemes in semitic. *Word*, 5, p. 162-181, 1950.

KEYSER, J.; STEVENS, K. Enhancement and overlap in the speech chain. *Language 82(1)*, p. 33-63, 2006.

MADDIESON, I. The structure of segment sequences. *UCLA Working Papers in Linguistics*, v. 83, p. 1-7, 1993.

OLIVEIRA, L. C. F. *Léxico, Alofonia e Percepção de Fala na Fonologia Articulatória.* 2003. Dissertação (Mestrado) – LAFAPE, IEL, UNICAMP, Campinas.

OLIVEIRA, M. A. A controvérsia neogramática reconsiderada. In: ALBANO, E. C.; COUDRY, M. I.; POSSENTI, S.; ALKMIM, T. M. (Org.). *Saudades da Língua: a Lingüística e os 25 Anos do Instituto de Estudos da Linguagem da UNICAMP*. Campinas, Mercado de Letras, p. 605-620.

PIERREHUMBERT, J. Probabilistic phonology: discrimination and robustness. In: BOD, R.; HAY, J.; JANNEDY, S. (Org.). *Probabilistic Linguistics*. Cambridge, Mass.: MIT Press, 2003. p. 177-228.

PIERREHUMBERT, J. Syllable structure and word structure. In: KEATING, P. (Org.). *Papers in Laboratory Phonology III*. Cambridge: Cambridge University Press, 1994. p. 168-190.

SILVA, A. H. P. Caracterização acústica de [R], [r], [l] e [L] nos dados de um informante paulistano. *Cadernos de Estudos Lingüísticos 37*, p. 51-68, 1999.

SILVA, A. H. P. *As fronteiras entre a fonética e a fonologia e a alofonia dos róticos iniciais em PB: dados de dois informantes do sul do país.* 2002. Tese (Doutorado) – Lafape, UNICAMP, Campinas, 2002.

SILVA, A. H. P.; MOREIRA, A.; VILLA, M. H.; AQUINO, P. Codificação fonológica informatizada do Minidicionário Aurélio: um banco de dados para o estudo da fonologia portuguesa. *Estudos Lingüísticos XXIII: Anais de Seminários do GEL*, v. 2, 1994. p. 1321-1327.

SPROAT, R.; FUJIMURA, O. Allophonic variation of English /l/ and its implication for phonetic implementation. *Journal of Phonetics* 21, p. 291-311, 1993.

TRAUNMÜLLER, H. Coarticulatory effects of consonants on vowels and their reflection in perception. *Proceedings from the XIIth Swedish Phonetics Conference*, Dept. of Linguistics, Göteborg University, 1999. p. 141 - 144.

TRAUNMÜLLER, H. Size and physiological effort in the production of signed and spoken utterances. *Lund University Working Papers, Department of Linguistics,* 49, p. 164-167, 2001.

VITEVITCH, M.; LUCE, P.A. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language 40*, p. 374-408, 1999.

ZIPF, G. K. *Human Behavior and the Principle of Least Effort.* Nova Iorque: Hafner, 1949.