O PAPEL DA ESTATÍSTICA E DE MÉTODOS QUANTITATIVOS APLICADOS À (SOCIO)LINGUÍSTICA

Maria Guadelupe Dourado Rabello [1]
Leônidas José da Silva Jr.[2]

RESUMO

No meio acadêmico, os docentes e discentes procuram "separar" áreas como humanidades, ciências da natureza, ciências da saúde dentre outras. Na área de linguística, por exemplo, os experimentos científicos precisam de validação e, na grande maioria das vezes, esta envolve procedimentos estatísticos pouco amigáveis a pesquisadores das áreas de humanidades. Tendo em vista a dificuldade de muitos linguistas ou estudiosos da área, em estruturar, quantificar, analisar, interpretar e reportar os dados de suas pesquisas, o presente artigo tem como objetivo mostrar como a matemática aplicada, em especial, por meio da Estatística, pode estar inserida na linguística em uma interface entre a linguagem matemática e a língua natural. Para o referencial teórico, utilizamos trabalhos como os de Labov (1966, 2008), Sankoff (2001) e Guy (2007) quando da formalização, estruturação e protocolos estatísticos adotados na pesquisa em (sócio)linguística. Para metodologia, realizamos uma análise em um *corpus* de L2 (cf. Silva Jr., 2009) para análise da fricativa interdental não-vozeada produzida por brasileiros a fim e realizamos um teste de qui-quadrado. Os resultados apontam uma forte correlação entre grau de escolaridade e produção do segmento-alvo. Concluímos que, sem uma análise estatística com um protocolo sistematizado, as descrições *por si*, podem levar a erros de interpretação dos resultados.

Palavras-chave: Métodos quantitativos, Sociolinguística, Estatística, Matemática aplicada.

INTRODUÇÃO

Ao longo de nossa experiência e de relatos de docentes nos últimos anos em sala de aula na rede de ensino público da educação básica, bem como da superior, não raro nos deparamos com dificuldades mais acentuadas no ensino e abstração de estatística e probabilidade; ferramentas matemáticas essenciais para formalização de protocolos de pesquisas científicas em qualquer área da ciência. Em se tratando das ciências humanas e levando-se em conta que os profissionais que atuam nestas áreas têm um distanciamento natural das ciências exatas por diversos fatores (dificuldades de aprendizado do tempo de escola, dedicação à sua área de estudo, etc) e que em um dado momento estes profissionais precisam fazer uso destas ferramentas para que suas pesquisas tenham

[1] Mestranda em Ciências da Linguagem pelo Programa de Pós-graduação em Ciências da Linguagem da Universidade Católica de Pernambuco (PPGCL/UNICAP/CAPES); Especialista em Ensino da Matemática pela Universdade Federal de Pernambco - PE, guadelupedr@gmail.com;
[2] Doutor em Linguística pela Universidade Federal da Paraíba - PB; Pós-doutorado em Fonética experimental pela Universidade Estadual de Campinas (UNICAMP/CNPq) - SP, leondas.silvajr@gmail.com.

credibilidade, faz-se necessário uma (re)tomada e/ou uma (re)entrada em conteúdos outrora estudados todavia, sem qualquer aplicabilidade.

Além disso, a pouca familiaridade que os professores de matemática têm nessa área, já que só recentemente as licenciaturas em matemática apresentam essa disciplina (muitas vezes como optativa) – assim conceitos como aleatoriedade, incerteza e variabilidade nem sempre são enfatizados e discutidos ao longo do conteúdo correspondente como afirmam Oliveira & Cordani (2016). Os autores ainda expõem que os livros didáticos muitas vezes só apresentam uma abordagem instrumental dos conceitos de probabilidade e quando o fazem, geralmente reportam-se à análise combinatória (ferramenta para cálculo), impedindo uma discussão mais ampla de análise de dados e da importância da probabilidade nas análises estatísticas como, por exemplo, calcular a probabilidade de um determinado fenômeno linguístico ocorrer em uma dada comunidade de fala e como interpretá-lo.

Muitas áreas relacionadas à sociolinguística, nas quais os dados quantitativos desempenham um papel, viram a aplicação de métodos estatísticos, tanto tradicionais (desenho experimental, amostragem, estimativa, teste de hipóteses), quanto heurísticos (agrupamento e escalonamento de uma quantidade incerta de eventos), comuns em áreas como a Psicologia Social; Sociometria; Pesquisas de opinião & atitude (pesquisas de intenção de voto, por exemplo) que se utilizam de técnicas de escalonamento bi e multidimensional.

Segundo Sankoff (2001), é a partir da Teoria da Variação, no entanto, onde as questões sociais tornam-se mais intimamente ligadas às questões fonéticas (variação de sotaque e aspectos da fala) e questões gramaticais (fonológicas, morfológicas e sintáticas), em que um protocolo sociolinguístico especificamente para análise estatística foi desenvolvido e amplamente adotado.

A linguística (ainda) é uma das únicas entre as disciplinas científicas, que um número significativo de seus pesquisadores não exige e nem faz uso de metodologia estatística e seus resultados não são limitados por critérios estatísticos de validade. Sankoff (2001) afirma que, os linguistas tradicionalmente concordavam que a estrutura gramatical de uma língua consistia, em grande parte, de entidades ou categorias discretas, cujas relações e restrições de co-ocorrência eram de natureza qualitativa e compartilhadas por todos os falantes da comunidade de fala. Essas estruturas podem então ser deduzidas analisando e comparando enunciados suscitados ou intuídos por qualquer falante nativo

da língua (por exemplo, linguistas que servem como sua própria fonte de dados), sem necessidade de qualquer aparato estatístico.

É apenas a partir do trabalho inovador de William Labov, no final dos anos 1960 (Labov, 1969), que de fato passa a haver uma formalização científica na montagem de desenhos experimentais e de protocolos estatísticos para investigar questões de interesse central da teoria linguística. Em seu trabalho, o autor examina, do ponto de vista da fonologia gerativa, a posição da cópula do verbo de ligação "*be*" (ser/estar) na fala do inglês não-padrão afro-americano *(Non-standard Negro English*) e conclui que de fato, há o apagamento desta cópula e que suas inferências fazem todo sentido visto que outras línguas como húngaro, hebraico e crioulo-francês do Caribe apresentam comportamento gramatical semelhante.

Com este trabalho, Labov para a comunidade linguística a ideia de que duas ou mais articulações distintas de uma determinada forma fonológica pode ocorrer na mesma palavra ou afixo, nos mesmos contextos, sem afetar o significado referencial de um item lexical ou mesmo a função sintática de um determinado afixo ou partícula. É possível prever qual forma ocorrerá em um determinado momento no tempo a partir de um modelo probabilístico, pelo qual os efeitos do contexto linguístico e extralinguístico podem ser determinados com precisão, todavia, o resultado desta análise continua sendo apenas uma probabilidade. A escolha da forma sempre contém um componente do puro acaso, embora isso seja precisamente delimitado.

Os modelos formais (a Linguística Formal) da teoria gramatical têm estruturas discretas de natureza algébrica, algorítmica e/ou lógica. Tais estruturas geralmente envolvem conjuntos de dois ou mais componentes alternantes, como sinônimos, paráfrases ou alofones, que o pesquisador pode determinar estar executando funções linguísticas idênticas ou semelhantes. Ao permitir um certo grau de aleatoriedade na escolha entre esses suplentes, os formalismos gramaticais são convertidos em modelos probabilísticos de desempenho linguístico suscetíveis ao estudo estatístico (CEDERGREN & SANKOFF, 1974).

Desta forma, o presente artigo tem como objetivo mostrar como a matemática aplicada, em especial, por meio da Estatística, pode estar inserida na linguística e, desta forma, facilitar o trabalho dos estudiosos e pesquisadores mostrando modelos matemático-estatísticos de forma prática de serem utilizados com uma interface entre a linguagem matemática e a língua natural.

A justificativa para o presente trabalho se dá tendo em vista a dificuldade de muitos linguistas ou estudiosos da área, em estruturar, quantificar, analisar, interpretar e reportar os dados de suas pesquisas.

Para Metodologia, analisamos a produção/apagamento do som da fricativa interdental não-vozeada (conhecida como o som do "*th*"). Para isso, utilizamos o corpus L2BRA_VOWELS (Silva Jr., 2009). Verificamos as ocorrências de apagamento/produção nos seguintes fatores: alunos do *Ensino fundamental*, *Ensino médio* e *Ensino superior* deste segmento-alvo e, assim como, Labov (1966), realizamos um teste de qui-quadrado.

Nossos resultados apontaram que alunos do Ensino superior produzem de forma robusta o som interdental quando falam inglês e que, com isso, é possível verificar uma significativa associação entre este nível e os demais fatores da variável categórica em estudo (*grau de escolaridade*). Em contrapartida, os alunos da educação básica não apresentaram diferença quanto à produção. Ambos não realizaram o segmento de modo consistente. Com essas observações e, mesmo que de modo preliminar, podemos concluir que o fator escolaridade foi preponderante para a realização de segmentos na L2.

Linguística, matemática e suas interfaces

Ainda na fase estudantil enquanto para lguns alunos no ensino regular, a Matemática é vista como uma disciplina "difícil", pois eles sempre questionam a sua pouca aplicabilidade, para outros ela "encanta". De acordo com Fayol (2012), aproximadamente 20% das crianças e adolescentes desenvolvem sentimentos negativos pela Matemática, que vão da ansiedade à fobia. No início do século XX, trabalhos acerca do *numeramento* matemático começam a ser mais explorados, sendo o *numeramento* a capacidade individual de formular, empregar e interpretar a matemática em uma variedade de contextos.

A BNCC vai além quando aciona um dispositivo em que determina o desenvolvimento do letramento matemático, que é o *numeramento*, definido como as competências e habilidades de raciocinar, representar, comunicar e argumentar matematicamente, de modo a favorecer o estabelecimento de conjecturas, a formulação e a resolução de problemas em uma variedade de contextos (BRASIL, 2018, p. 266). É por esse motivo que, quando um professor mostra ao aluno, desde cedo, que a Matemática é uma disciplina que pode estar inserida em diversos contextos e em diversas disciplinas,

pode ser um passo para que ele tenha interesse, empolgação e mostre-se disposto a aprender.

Ao longo da história moderna de sua ciência, a Linguística Formal (LF) (Fonética, Fonologia, Morfologia, Sintaxe e Semântica Formal) de modo não incomum, tem feito uso da Matemática, da Estatística e da Lógica como uma ferramenta para investigar, postular e quantificar aspectos da língua a partir de modelos experimentais além de refutar propostas teóricas em função de resultados empíricos, pois de modo geral, a LF entende linguagem por si como um sistema matemático baseado em regras do Cálculo e da Estatística (Fonética), bem como em regras da Lógica (as demais áreas citadas acima). Seus achados têm revelado cada vez mais aspectos importantes sobre a linguagem bem como, propósitos de como ela pode ser usada.

Assim, pesquisadores recorrem à Estatística para chegar a um resultado final da pesquisa sociolinguística, pois, segundo Scherre e Naro (2003), a Estatística entra nesse processo para "revelar tendências e correlações inerentes na massa de dados linguísticos, e validá-las, dentro de um determinado grau de certeza". Assim, a Estatística ajuda o pesquisador-sociolinguista a quantificar, resumir e manipular os grandes dados coletados. Portanto, é através da quantificação que o pesquisador confere suas especulações, realiza análises, interpreta os resultados obtidos e afirma o resultado da sua pesquisa. É muito comum o uso de tabelas e gráficos para ilustrar melhor a variação estudada, sendo possível através de tal processo, conferir se há na Língua uma variação estável ou uma mudança em curso.

Os modelos matemático-estatísticos foram e são utilizados por estudiosos e pesquisadores da área de Linguística, pois de acordo com Sell e Gonçalves (2011), a Sociolinguística Variacionista surgiu na década de 1960 a partir dos estudos de William Labov, nos Estados Unidos, e desde então vem desenvolvendo pesquisas sobre variação e mudança linguísticas nas mais variadas línguas, sendo esse modelo de análise linguística também conhecido como sociolinguística quantitativa, já que trabalha com o tratamento estatístico e sistemático dos dados coletados.

Uma das principais ideias de Labov é de que a variação é inerente à linguística e também necessária para o funcionamento de uma língua. A comunidade de fala é o ponto de partida e o objeto de observação da análise sociolinguística. Para Labov (1972), a definição de língua não pode estar desvinculada do social que exerce uma função comunicativa em grupos sociais/culturais. Logo, em uma comunidade de fala, os indivíduos devem compartilhar atitudes e normas que tenham características linguísticas

semelhantes, facilitando assim as pesquisas com o enfoque amplo sobre o fenômeno de variação.

O conceito-chave subjacente à sociolinguística variacionista é a 'variável linguística'. Vamos apresentar aqui na Tabela 1, um exemplo com base no trabalho de Labov (1968) em que observa-se a distribuição de ocorrências do verbo copular 'be' (ser/estar) do inglês falado, que ocorre como contração no inglês padrão (*Standard English* - SE) e apagamento no inglês não-padrão afro americano (*(Non-standard Negro English* – NNE).

FORMA REGULAR	SE	NNE	TRADUÇÃO
*John **is** a doctor*	*John**'s** a doctor*	*John a doctor*	John é médico
*We **are** there*	*We**'re** there*	*We there*	Estamos lá
*I **am** at home*	*I**'m** at home*	*I at home*	Estou em casa

Tabela 1: Variáveis linguísticas para a cópula "*be*" propostas por Labov (1969).

Outro exemplo do uso de variável linguística (o qual nos deteremos mais detalhadamente nas próximas seções) envolve a pronúncia das fricativas interdentais (o som do "*th*"), como em '*this* – [ð]*is*' (isto), vozeada, e '*think* – [θ]*ink*' (pensar), não-vozeada, que também são pronunciadas, pelo menos ocasionalmente, pelos falantes de grande parte das variedades de inglês menos prestigiadas e inglês como língua estrangeira (L2) como oclusivas "*dis**" e "*tink**".

De acordo com Bagno *apud* (LABOV, 2008), embora o impacto dos trabalhos de Labov em estudos da linguagem seja amplamente reconhecido, o seu conceito de "social" vem sendo criticado por estudiosos filiados a outras correntes teóricas, como a Análise do Discurso, a Sociologia da Linguagem, a Antropologia Linguística e a Sociolinguística Interacional. Bagno ainda menciona que é inegável que a Sociolinguística Variacionista tem fornecido suporte empírico para o combate às construções que se apoiam nas diferenças linguísticas como pretexto para políticas de discriminação e exclusão social.

Para compreendermos os métodos que Labov utilizava, há necessidade de que se tenha alguma noção sobre métodos de investigação empírica. De acordo com Monteiro (2000), o variacionismo parte do pressuposto de que a heterogeneidade manifestada na fala pode ser analisada de forma que o pesquisador deve colher uma boa soma de dados em uma comunidade e que esses dados construirão o material que será submetido a análises estatísticas para a testagem da sua hipótese. O autor menciona ainda que a opção

pela pesquisa empírica se liga ao fato de que, sendo a sociolinguística uma ciência social, ela depende da observação do comportamento do homem.

De acordo com Guy (2007), toda pesquisa - dialetal, geográfica ou social - é inerentemente quantitativa, visto que essa metodologia inclui uso de tabelas e gráficos para apresentação de dados, teste de significância, confiabilidade e técnicas analíticas quantitativas. O autor ainda menciona que existem três fases principais no curso de qualquer análise quantitativa, são elas:

- *Coleta de dados*: Aqui, devemos observar aspectos como amostra e confiabilidade em que Guy defende que em estudos de comunidade de fala, onde se usam mais de um pesquisador, há prática de se usar testes de confiabilidade entre os pesquisadores para assegurar que todos estão aplicando o mesmo critério de análise;
- *Redução e apresentação de dados*: As técnicas para redução de dados mais utilizados, no que diz o autor, provêm da área de Estatística em que tais técnicas incluem medidas de tendências centrais como média, mediana e moda usando tabelas, gráficos ou mapas como os principais métodos de apresentação usados em pesquisa dialetal;
- *Interpretação e explicação de dados*: Neste ponto, Guy (op. cit) aponta que o objetivo final de qualquer estudo quantitativo em pesquisa dialetal é identificar e explicar fenômenos linguísticos e não somente produzir números, como por exemplo, medidas estatísticas para resumir os dados. Os fenômenos e sua natureza devem ser explicados *através* dos números. Estes (os números) mostram, de forma codificada em linguagem matemática, o que acontece com um determinado aspecto de variação fonética, por exemplo. Em outras palavras, os números representam uma espécie de "bússola", i.e., um caminho que pode explicar o comportamento linguística de uma dada comunidade de fala;

O autor conclui que é através da realização de análises quantitativas que é possível realizar estudos da variação linguística em uma dada comunidade. Quando fala-se em variação, esta deve ser compreendida como a alternância entre dois ou mais elementos linguísticos, que não pode ser adequadamente descritos e analisados em termos categóricos ou estritamente qualitativos. Assim, na sua concepção, um modelo quantitativo na sociolinguística variacionista vai existir quando tomamos um modelo de

teoria linguística que procura explicar as possibilidades linguísticas e os padrões quantitativos do uso dessas possibilidades através de um modelo matemático-estatístico.

Como exemplo de modelamento matemático-estatístico, temos o teste de Qui-quadrado (χ^2), em que Labov foi o pioneiro ao utilizá-lo na linguística em suas pesquisas/experimentos de cunho quantitativo.

O teste Qui-quadrado (χ^2)

De acordo com Guy (2007), a distribuição qui-quadrada ou (teste) Qui-quadrado (do inglês, *Chi-squared distribution/test*) é um procedimento útil para calcular a probabilidade de uma hipótese nula (H_0) ser verdadeira. Sua estatística é uma medida de divergência entre a distribuição dos dados e uma distribuição esperada dos dados em análise.

A técnica testa a independência ou determina a associação entre as variáveis categóricas. Por exemplo, se você tem uma tabela de dois *fatores* (por exemplo, periferia/bairro nobre) com os resultados pela variável *classe social*, a estatística de qui-quadrado pode ajudar a determinar se as pessoas que compram em um determinado *shopping center* é independente da *classe social* ou se há alguma associação entre classe social e compras neste *shopping center*. Se o valor-p (*p-value*) associado à estatística qui-quadrado for menor do que seu α selecionado (comumente, pesquisas em nas ciências humanas, p é estabelecido em 5%), o teste rejeita a hipótese nula de que as duas variáveis são independentes.

Este teste é uma das distribuições mais utilizadas em Estatística Inferencial e serve para avaliar quantitativamente a relação entre o resultado de um experimento e a distribuição esperada para o fenômeno. Isto é, ele nos fornece informações se os valores observados podem ser aceitos pela teoria em questão.

O teste de Qui-quadrado é calculado através da seguinte fórmula (Eq.) em Eq. (1) e explicada em notação linguística em Eq. (2):

Eq. (1)

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \qquad E = \frac{\prod LC}{N}$$

Onde:

- O_i= Frequência observada para cada classe;
- E_i = Frequência esperada para aquela classe;
- K = Número total de observações;

- N = total de amostras.

Eq. (2)

$$\chi^2 = \sum_{i=cada\ ocorrência}^{todas\ as\ ocorrências} \frac{(dados\ observados - \ dados\ esperados\ )^2}{dados\ esperados}, \ E = \frac{\sum linha\ x \sum coluna}{Total\ da\ amostra}$$

Análise linguístico-quantitativa com base em *amostras* de dados de fala: o trabalho de Labov (1966)

A linguagem modelada em termos probabilísticos é mais apropriadamente estudada pelos dados de fala natural: sequências fluentes sustentadas com base no encadeamento de enunciados sem previsibilidade de quando ou com que frequência o fenômeno linguístico em estudo ocorrerá no fluxo da conversa. Portanto, a *amostra* de dados de fala geralmente envolve relativamente poucos participantes (20 a 120), cuidadosamente escolhidos para representar a diversidade de comportamento linguístico dentro da comunidade em estudo, com um grande volume de material gravado de cada falante (SANKOFF, 2001).

Como exemplo de análises quantitativas de variações linguísticas em que a análise de *amostras* de fala são levadas em conta, temos uma pesquisa realizada por Labov que tornou-se conhecida, e até hoje se configura como estudo pioneiro e uma das referências mais importantes relacionadas aos fatores sociais em relação ao falar de uma localidade.

De acordo com Labov ([1966], 2008), a pesquisa foi realizada em 1966, na cidade de Nova Iorque em três lojas de departamento. Ele partiu da hipótese de que a realização do fonema seria determinada pelo ambiente socioeconômico em que o falante se encontra. Labov apresenta um estudo que procura compreender as variações fonológicas surgidas a partir da produção da consoante [r] em posição pós-vocálica, observando as condições sociais dos falantes de Nova Iorque, onde seleciona ambientes no qual desenvolverá suas observações e estes estabelecimentos apresentam diferentes aspectos (*variáveis*) sociais como por exemplo: em relação aos preços e às pessoas que os frequentam como *sexo*, *classe social*, *faixa etária* e *raça*. As lojas selecionadas por Labov foram: a *Saks* (*status* financeiro superior), a *Macy's* (*status* financeiro médio) e *S. Klein* (*status* financeiro mais baixo).

O método utilizado por ele segue um procedimento específico em que se aproximava do informante no papel de um cliente e pedia informações em relação ao local de sapatos femininos. A resposta geralmente era: *"fuurth floor"*. O entrevistador, no caso o próprio Labov, ainda se inclinava para frente e perguntava: "*excuse-me?*"

("como?"). Normalmente ele obtinha outro enunciado: *"fou**R**th floo**R**"* pronunciado no estilo mais monitorado. As *variáveis* identificadas pelo uso do 'r' foram as ocorrências *casuais* (*fou**r**th fou**r***, sem monitoramento) e as enfáticas (*fou**R**th fou**R***, com monitoramento). O pesquisador tomou nota das situações em que houve a ocorrência de africadas e de oclusivas para a consoante em momento final do vocábulo *fourth*, atentando-se também às variações não padronizadas da interdental '*th*' utilizadas pelo falante.

Labov fez 68 entrevistas na *Saks*, 125 na *Macy's* e 71 na *S. Klein*, totalizando 6 horas e 30 minutos de tempo, distribuídos entre os 264 falantes. Para cada ocorrência totalmente constritiva da variável, ele registrou (r-1); para nenhuma fonação, registrou (r-0). Para ilustração, o autor, em seu livro, faz tabelas e gráficos bem elaborados para fazer comparações entre as características da fala das referidas lojas, onde nos deteremos a seguir na ilustração de apenas um dos gráficos, a fim de fazermos essa comparação no uso de (r) pelos funcionários de *Saks*, *Macy's* e *S. Klein*.

As Tabelas 2 e 3 apresentam os resultados da referida pesquisa (Labov (1966, 2008). A Tabela 2, apresenta os números absolutos da produção e apagamento do /r/ por loja de departamento, enquanto que a Tabela 3, apresenta o percentual da produção e apagamento do /r/ para loja de departamento.

Apagamento (*r-0*) Produção (*r-1*)	Lojas de departamentos			TOTAL
	S. Klein	Macy's	Saks	
r-0	195	211	93	499
r-1	21	125	85	231
TOTAL	216	336	178	730

Tabela 2: Tabela de contingência 2x3 da distribuição absoluta da produção/apagamento do /r/ por loja de departamentos em Nova Iorque (Labov, [1966], 2008).

Apagamento (*r-0*) Produção (*r-1*)	Lojas de departamentos		
	S. Klein (%)	Macy's (%)	Saks (%)
r-0	90.2	62.8	52.2
r-1	9.8	37.2	47.8

Tabela 3: Tabela da distribuição percentual (%) da da produção/apagamento do /r/ por loja de departamentos em Nova Iorque (Labov, [1966], 2008).

Com os valores apresentados na Tabela 3, observemos no Gráfico 1, a descrição dos dados de produção/apagamento do /r/ da referida pesquisa de Labov.

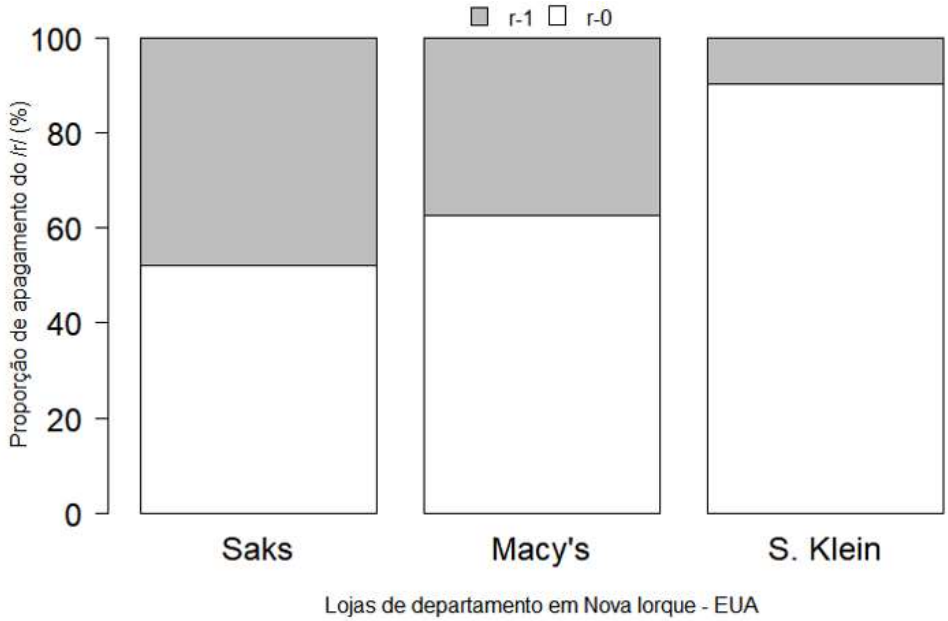

Gráfico 1: Apagamento e produção do /r/ pós-vocálico (Labov, 1966).

Vejamos na Tabela 4 os resultados dos testes de qui-quadrado a partir dos dados da Tabela 2:

LOJAS DE DEPARTAMENTO	VALOR χ^2	P-VALOR
Saks; Macy's; S.Klein	35,7	<0,01***
Saks; Macy's	2,28	=**0,13**
Saks; S.Klein	35,3	<0,01***
Macy's; S.Klein	21,0	<0,01***

Tabela 4: Resultados dos testes de Qui-quadrado do experimento de Labov (1966)

A partir do trabalho de Labov, podemos tirar conclusões a respeito desse fenômeno linguístico. Uma dessas conclusões é que, através da verificação do apagamento/manutenção do /r/, é possível mapear probabilisticamente o perfil socioeconômico dos clientes que frequentam as lojas da pesquisa: *Saks, Macy's* e *S. Klein*.

É possível detectar que o perfil de clientes que frequentam a *Saks* e *Macy's*, tem comportamento muito próximo, não por percentuais, mas pelos resultados da análise do Qui-quadrado, até porque, como foi dito no início do problema, a loja *Saks* é frequentada por pessoas que têm um *status* financeiro superior, enquanto que na *Macy's*, a frequência é de pessoas com *status* financeiro médio e o teste do qui-quadrado vai comprovar que não existem diferenças significativas entre os clientes que frequentam as duas lojas tomando como base o fenômeno fonético-fonológico do apagamento do /r/realizado pelos vendedores dessas duas lojas.

Já na loja mais popular (*S. Klein*) o apagamento do /r/ é significativamente maior do que nas demais lojas. Na loja mais popular, há uma preocupação consideravelmente menor quanto á utilização de uma linguagem de prestígio, tomando como base o apagamento do /r/. Confirma-se assim a hipótese de Labov, de que a realização do fone seria determinada pelo ambiente socioeconômico em que o falante se encontra.

É claro que isso é uma conclusão preliminar, todavia é um dos indícios apontados para o tratamento de questões socioeconômicas, as quais, na sociedade norte-americana, também estão correlacionadas às questões étnicas. Podemos observar como as análises estatísticas podem fazer com que possamos tomar decisões cotidianas a partir de fenômenos linguísticos e este fato aponta uma para uma relação positiva entre a matemática (por meio da estatística), e como ela pode auxiliar nas análises de *corpora* diversos; seja em língua materna ou em língua estrangeira.

Assim, a respeito desta pesquisa, Labov ([1966], 2008) menciona que entrevistas rápidas e anônimas podem ser uma fonte valiosa de informações sobre a estrutura sociolinguística de uma comunidade de fala, pois toma como seu primeiro objetivo, a língua usada por pessoas comuns em seus afazeres cotidianos, embora menciona também que há caminhos por onde se possa ampliar e melhorar esses métodos, como por exemplo, o autor cita que os dados poderiam ter sido gravados para que se obtivesse uma melhor transcrição dos casos duvidosos.

Reportando os resultados estatísticos do teste de Qui-quadrado em trabalhos acadêmicos

Uma vez que o(a) pesquisador(a) realizou as análises e interpretações acerca dos dados, faz-se necessário que os resultados sejam reportados de modo sistemático em publicações, tais como; periódicos, anais de congressos, etc. e/ou apresentações de trabalhos. Oushiro (2017, p. 108-109) orienta que a notação convencional (a matemática) é “lida” da seguinte forma em um exemplo hipotético: $\chi^2 = 74{,}14$ (2), $p<0{,}01$ "*Qui-quadrado igual a 74,14, com dois graus de liberdade e p menor do que 0.01*". A autora ainda explica cada uma das etapas: “**Qui**" é representado pela letra *chi* (χ) e o “**quadrado**” é o número ‘2’ sobrescrito.

A Tabela 5 apresenta, de forma mnemônica (por cores), as notações matemática e linguística para reportar o resultado de qui-quadrado nos experimentos.

Notação matemática	*Notação linguística*
$\chi^2 = 52{,}15\ (2),\ p<0{,}01$	***Qui-quadrado*** igual a **52,15**, **com dois graus de liberdade** e ***p*** **menor do que 0.01**

Tabela 5: Como reportar o resultado do teste de Qui-quadrado (*notação matemática* à esquerda) e como ler este resultado (*notação linguística* à direita)

Até aqui, vimos como matemática aplicada (via estatística) está presente nos fenômenos relacionados à linguagem; seja do ponto de vista fonético-fonológico, morfossintático, seja do ponto de vista semântico-formal. Apresentaremos na próxima seção, um procedimento estatístico semelhante ao de Labov, todavia, com dados de inglês como L2 produzido por falantes brasileiros.

METODOLOGIA

Tomando como ideia o experimento de Labov com o apagamento do /r/ e utilizamos o *corpus* da pesquisa de Silva Jr. (2009). Esta pesquisa se debruça sobre a análise da produção de vogais do inglês como língua estrangeira (L2) por alunos brasileiros com estratificação social pela *escolaridade* (ensino fundamental (**EF**), ensino médio (**EM**) e ensino superior (**ES**)) *tipo de escola* (pública, particular e universidade), *faixa etária* (**A**: 11-14 anos; **B**: 15-17 anos; **C**: acima de 17 anos). A coleta de dados foi realizada em escolas e universidade na cidade de Recife-PE O *corpus* utilizado na pesquisa foi o *L2BRA_VOWELS* do autor: <https://phonetics_prosodyl2.wixsite.com/l2bra_vowels>[3].

No referido *corpus*, há um texto no qual os participantes leram a seguinte passagem: "(...) *and I was about to finish. It was the* ***fourth*** *and last part of the treatment and the* ***therapy*** *was no longer so bad at that point* (...)". ("[...] e eu estava prestes a terminar. Foi a **quarta** e última parte do tratamento e a **terapia** já não estava tão ruim àquela altura [...]" (tradução nossa).

Escolhemos apenas a variável *escolaridade* para manter a condição de igualdade de variáveis com o experimento de Labov aqui utilizado.

Assim, retiramos para nossa análise a produção das palavras "*four****TH***" (quarta) e "***TH****erapy*" (terapia) com o intuito de verificar o apagamento/produção do som do "*th*" –

[3] O *corpus* está em fase desenvolvimento. Dados de novas coletas e a inserção de novas variáveis como: **nível de proficiência na L2**, **frequência de uso de inglês pela internet**, **frequência de uso de videogames online e off-line** serão em breve disponibilizadas no *corpus*.

fricativa interdental desvozeada representada pelo fonema /θ/. Este som do inglês é pronunciado com certa dificuldade pelos brasileiros por ele não estar presente em nosso inventário fonológico (cf. Silva Jr., 2019 para detalhamento da literatura fonética e sociolinguística que aborda este problema). Uma vez contabilizadas as produções, realizamos os mesmos procedimentos que o fizemos com o experimento de Labov ([1966], 2008).

O tratamento estatístico de nossos dados foi realizado em Linguagem R (R Core Team, 2020) disponível em:<https://cran.r-project.org/>.

Como dito na seção anterior, utilizamos o Teste Qui-quadrado para análise estatística dos nossos dados. Como vimos na seção anterior, esta técnica é usada para verificar se há uma associação entre as variáveis (de linha e coluna - cf. Tabela 6) (cf. Lowie & Seton, 2013; Triola, 2014) que compõem uma tabela de contingência construída a partir dos dados da amostra. Um valor de significância (alfa) de 5% foi utilizado para verificar se há mudanças significativas (ou não) na produção do segmento-alvo provocadas pelo *grau de escolaridade*. Um valor de 5% foi estabelecido para *alfa*. Se *alfa* < 5% (p<0,05), o *grau de escolaridade* influencia na *produção*/*apagamento* do segmento-alvo, ou seja, há associação entre o fenômeno em xeque e o fator *escolaridade*.

Com o propósito de tornar nossa análise quantitativa mais didática, vejamos na arquitetura da Tabela 6, como é feito o procedimento para identificar as linhas e colunas, previamente mencionadas nesta seção, bem como, os valores em porcentagem que utilizamos na aplicação do teste de Qui-quadrado (cf. Tabela 5 na próxima seção para detalhamento dos resultados):

Linhas	**Colunas**			
	E. Fund.	E. Med.	E. Sup.	TOTAL
th-0	30	34	11	75
th-1	10	6	29	45
TOTAL	40	40	40	120

Tabela 6: Tabela de contingência 2x3 com a produção absoluta dos dados e a representação das variáveis de linha (apagamento, *th-0* e produção, *th-1*) e de coluna (ensino *fundamental*, *médio* e *superior*).

A partir da equação matemática detalhada em Eq. (1) por notação convencional e Eq. (2) por notação matemático-linguística do teste qui-quadrado, em Eq. (3), vejamos como ela se aplica a nossos dados:

Eq. (3)

$$\chi^2 = \sum_{i=cad}^{todas\ as\ ocorrências} \sum_{ocorrência} \frac{(nossos\ dados\ 'th' - \ chances\ de\ ocorrer\ 'th')^2}{chances\ de\ ocorrer\ 'th'},$$

em que, $chances\ de\ de\ ocorrer\ 'th' = \frac{\text{SOMA das linhas } x \text{ SOMA das colunas}}{\text{total dos nossos dados}}$

RESULTADOS E DISCUSSÃO

APAGAMENTO (*th*-0)/PRODUÇÂO (*th*-1)	Grau de escolaridade			TOTAL
	EF	EM	SUP	
th-0	30	34	11	75
th-1	10	6	29	45
TOTAL	40	40	40	120

Tabela 7: Tabela de contingência 2x3 de distribuição absoluta para produção/apagamento do /θ/ por Grau de escolaridade.

APAGAMENTO (*th*-0)/PRODUÇÂO (*th*-1)	Grau de escolaridade		
	EF (%)	EM (%)	ES (%)
th-0	75	85	27.5
th-1	25	15	72.5

Tabela 8: Tabela de da distribuição percentual (%) para produção/apagamento do /θ/ por Grau de escolaridade.

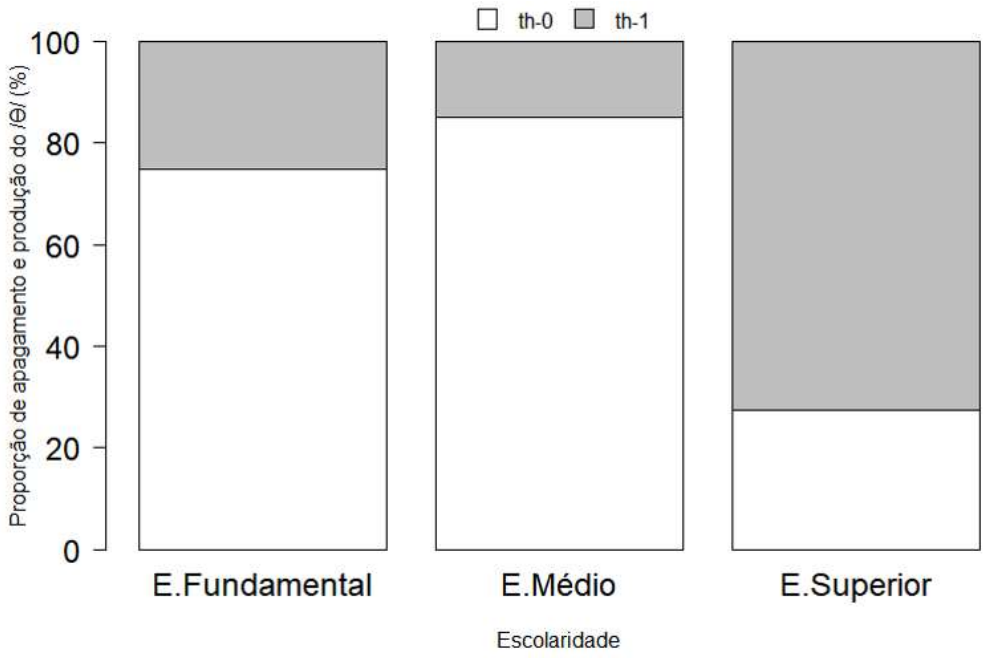

Gráfico 2: Proporção de apagamento e produção do /θ/ (corpus L2BRA_VOWELS - Silva Jr., 2009).

Vejamos na Tabela 9 os resultados dos testes de qui-quadrado a partir dos dados da Tabela 7:

ESCOLARIDADE	VALOR χ^2	P-VALOR
EF; EM; ES	32,21	<0,001***
EF; EM	1,25	**>0,26**
EF; ES	18,06	<0,001***
EM; ES	26,87	<0,001***

Tabela 9: Resultados dos testes de Qui-quadrado (χ^2) e valor de *p* (P-VALOR) para apagamento/produção (*th*-0/*th*-1, respectivamente) da fricativa interdental não-vozeada [θ] a partir da combinação das *categorias* de ESCOLARIDADE: ensino fundamental (EF), ensino médio (EM) e ensino superior (ES).

Os resultados apontam que, os alunos do ensino superior, produzem significativamente mais o som do "th" nas palavras por nós estabelecidas do que os alunos do ensino fundamental e médio. Desta forma, podemos afirmar que o nível de escolaridade tem considerável influência sobre a produção do inglês como L2 no que tange este som em específico. O teste de qui-quadrado também revela que; entre alunos da educação básica (ensino fundamental e médio) não há diferenças significativas quanto à produção embora, percentualmente, a produção do "th" dos alunos do ensino fundamental é ligeiramente maior do que médio. Entre os níveis EF-ES e EM-ES há diferenças significativas quanto a produção do fenômeno.

Seria precoce afirmar que os alunos na educação básica apagam mais o som do "th" que uma das limitações do corpus é não apontar o nível de proficiência do falante bem como, outras variáveis como o fato de assistir programas apenas em inglês ou mesmo verificar a frequência de uso da internet. Vale ressaltar que, esses dados datam de um período em que não havia *smartphones* com aplicativos para se aprender ou aperfeiçoar a pronúncia na L2-alvo e o acesso à internet era bem mais limitado. Esta pesquisa escolheu aleatoriamente alunos da educação básica de escolas públicas e particulares em condições de igualdade numérica.

Vimos então que a análise de um fenômeno linguístico a partir de matemática estatístico-probabilística é possível traçar perfis de tendências no campo científico; seja nas ciências de saúde, da tecnologia ou sociais. A ferramenta matemática é nossa aliada quando pensamos em uma relação determinística e dedutiva tendo como marco zero, modelos que nos auxiliam a fazer estimativas de uma dada situação.

De algumas décadas para nossa atualidade, os programas computacionais como (ambiente) **R**, *Stata*, *Microsoft Excel*, *SPSS* e tantos outros têm realizado com alta

eficiência a tarefa de calcular testes e outros procedimentos estatísticos. Atualmente, também é possível realiza-los a partir de dispositivos móveis, tais como, *smartphones* e *tablets* ou mesmo por manipuladores virtuais online. Mas, seria interessante mostrar como este cálculo acontece para que possamos quebrar paradigmas quanto ao suposto "alto grau de dificuldade" que um determinado teste estatístico venha a nos imputar.

Vejamos em Eq. (4) a fórmula/equação de qui-quadrado – desde seu momento abstrato à sua aplicação – a partir de sua decomposição em favor de nossos dados localizados na Tabela 7.

Eq. (4)

$$\boldsymbol{\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}}, em\ que\ E = \frac{\prod LC}{N} \rightarrow E = \frac{\sum linha\ x \sum coluna}{Total\ da\ amostra}$$

Onde:

- O_i= Frequência observada para cada *classe* (escolaridade);
- E_i = Frequência esperada para aquela *classe* (escolaridade);
- K = Número total de observações;
- N = total de amostras.

$$\chi^2 = \sum_{i=cada\ ocorrência}^{todas\ as\ ocorrências} \frac{(nossos\ dados\ 'th' - chances\ de\ ocorrer\ 'th')^2}{chances\ de\ ocorrer\ 'th'},$$

$$\text{em que, } chances\ de\ de\ ocorrer\ 'th' = \frac{\text{SOMA das } \mathbf{linhas}\ x\ \text{SOMA das } \mathbf{colunas}}{\text{total dos nossos dados}}$$

Em seguida, a Frequência esperada (**E**) da produção/apagamento para uma das classes na Eq. (5):

Eq. (5)

$$E(th0/th1, EF, EM, ES) = \frac{(th0)(EF) + (th0)(EM) + (th0)(ES)}{Total\ da\ amostra} + \frac{(th1)(EF) + th1(EM) + (th1)(ES)}{Total\ da\ amostra}$$

Uma vez decomposta a Eq (4), vejamos sua aplicação prática na Eq. (6).

Eq. (6)

$$\chi^2 = \frac{(30-25)^2}{25} + \frac{(34-25)^2}{25} + \frac{(11-25)^2}{25} + \frac{(10-15)^2}{15} + \frac{(6-15)^2}{15} + \frac{(29-15)^2}{15},$$

$$\chi^2 = \frac{25}{25} + \frac{81}{25} + \frac{196}{25} + \frac{25}{15} + \frac{81}{15} + \frac{196}{15} = \frac{25+81+196}{25} + \frac{25+81+196}{15} = \frac{302}{25} + \frac{302}{15}$$

$$\chi^2 = \mathbf{32.21}$$

A Eq. (6), foi utilizada para calcular o valor do qui-quadrado entre os três fatores do grau de escolaridade, ou seja, todas as linhas e colunas da Tabela 7. Seu resultado figura na primeira linha da Tabela 9. Para calcular os demais valores, repita o procedimento para cada linha e coluna da Tabela 7 que corresponda sua análise.

CONSIDERAÇÕES FINAIS

Neste trabalho, mostramos a importância da aplicabilidade de modelos e métodos matemático-estatísticos em pesquisas na área de sociolinguística de L2. Além disso, apresentamos como tais procedimentos quantitativos permearam de forma primordial e preencheram lacunas nas áreas de conhecimento envolvendo as ciências humanas no intuito de protocolar, sistematizar e seguir um determinado protocolo experimental.

Apresentamos aqui os trabalhos seminais de Labov, pioneiro em utilizar métodos quantitativos para se fazer compreender em suas análises linguísticas e resultados de seus estudos e pesquisas. Procuramos mostrar esses modelos de uma forma prática com a intenção de contribuir na formação de pesquisadores que precisam realizar análises quantitativas em pesquisas de caráter experimental.

Os resultados de nossa pesquisa apontam forte associação entre o grau de escolaridade e a produção e/ou apagamento do som do ‘th’. Essas considerações indicam que o estudante universitário procura buscar mais questões relaciionadas à pronúncia da L2 por alguns motivos distintos como, o aumento do interesse pelo idioma em função de poder tentar um bom emprego, a necessidade por causa da cobrança do meio acadêmico, viagens ao exterior em programas vinculados à sua instituição de ensino, dentre outros. O fato é que: as práticas de ensino de inglês como L2 na educação básica devem ser revisitadas; desde o investimento na formação continuada de docentes de L2, a condições e materiais compatíveis à realidade do aluno.

Deste feito, podemos concluir que modelos matemáticos quando aplicados, atuam não apenas no foco da pesquisa, como vimos nos dados aqui analisados, mas podem ser usados para refletir realidades e práticas de ensino em salas de aula por professores/pesquisadores de L2.

Por fim, foi apenas possível inferir tais resultados e realizar uma interpretação embasada na realidade, por causa da análise estatística por nós utilizada. Do contrário, teríamos apenas conjuntos de dados descritos, no máximo, em porcentagem que, por

vezes, podem nos guiar a erros indesejáveis na pesquisa e mascarar a realidade pela ausência de argumentos interpretativos.

AGRADECIMENTOS

Agradecemos a concessão de bolsa à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), sob o nº. 88887.483123/2020-00, para a primeira autora e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), sob o nº. 150143/2018-4, para o segundo autor.

REFERÊNCIAS

BRASIL. *Base Nacional Comum Curricular: Ensino Médio*. Brasília: MEC/Secretaria de Educação Básica, 2018, pp. 527-546.

CEDERGREN, H.; SANKOFF, D. Variable rules: performance as a statistical reflection of competence. **Language** (50), p. 333–355, 1974.
FAYOL, M. *Numeramento: aquisição das competências matemáticas*. São Paulo: Parábola Editorial, 2012.

GUY, G; ZILLES, A. *Sociolinguística quantitativa*. São Paulo: Parábola Editorial, 2007.

LABOV, W. A estratificação social do (r) nas lojas de departamento na cidade de Nova York. [1966]. In: LABOV, W. *Padrões Sociolinguísticos*. Tradução Marcos Bagno. São Paulo: Parábola, 2008.

LABOV, W. *Sociolinguistic Patterns*. Pennsylvania: University of Pennsylvania Press, 1972.

LABOV, W. Contraction, deletion, and inherent variability of th eEnglish copula. **Language** (45), p. 715–762, 1969.

LOWIE, W.; SETON, B. *Essential Statistics for Applied Linguistics*. Ney York: Palgrave Macmillan, 2013.

MOLICA, M.C.; BRAGA, M.L. *Introdução à Sociolinguística e tratamento de Variação*. São Paulo: Contexto, 2003.

MONTEIRO, J. *Para compreender Labov*. Petrópolis: Editora Vozes, 2000.

OLIVEIRA, C.; CORDANI, L. Julgando sob incerteza: heurísticas e vieses e o ensino de probabilidade e estatística. **Educ. Matem. Pesq**. 18(3), p. 1265-1289, 2016.

OUSHIRO, L. *Introdução à Estatística para Linguistas*. vol 1, Unicamp, 2017. Disponível em: <https://rpubs.com/oushiro/iel>.

SANKOFF, D. Statistics in Sociolinguistics. *Concise Encyclopedia of Sociolinguistics*, Elsevier, p. 828-834, 2001.

SELL, F.; GONÇALVES, A. *Sociolinguística*. Indaial: Uniasselvi, 2011.

SILVA Jr. L. *Análise Acústica da Produção da Fricativa Interdental Desvozeada /Θ/ no Inglês como L2*, vol 1, UEPB, 2019. Disponível em: <https://sistemas.uepb.edu.br/epibic/# >.

SILVA Jr. L. *Erro de leitura das vogais do inglês americano como língua estrangeira pelos falantes do português do Brasil*. Dissertação (Mestrado) - Universidade Federal da Paraíba, João Pessoa, 2009.

TRIOLA, M. *Introdução à Estatística: Atualização da Tecnologia*. 11 ed. São Paulo: Ed. Saraiva, 2014.