Ferramentas metodológicas para análises (sócio)linguísticas

Rosane de Andrade Berlinck
Universidade Estadual Paulista (UNESP), Araraquara, São Paulo, Brasil
berlinck@fclar.unesp.br
http://orcid.org/0000-0003-3420-5541

Caroline Carnielli Biazolli
Universidade Estadual Paulista (UNESP), Araraquara, São Paulo, Brasil
caroline.biazolli@fclar.unesp.br
http://orcid.org/0000-0002-8578-8102

DOI: http://dx.doi.org/10.21165/el.v47i1.2047

Resumo
Uma das etapas mais relevantes em toda pesquisa científica é a escolha do aporte metodológico que viabilizará a sua realização. Essa escolha depende, naturalmente, do objeto de estudo e da abordagem teórica adotada. Muitos modelos teóricos vigentes nos estudos linguísticos investem em análises empíricas, enfrentando os desafios gerados pela manipulação de grandes volumes de dados. Em relação aos estudos (sócio)linguísticos, a metodologia da Teoria da Variação e Mudança Linguísticas se destaca pela centralidade que atribui à obtenção, análise e interpretação de dados. Objetivamos, neste artigo, apresentar as ferramentas metodológicas computacionais AntConc, Excel e Goldvarb X, que auxiliam no cumprimento de todas as etapas fundamentais para a concretização de estudos da variação e mudança linguísticas.
Palavras-chave: ferramentas metodológicas; variação e mudança linguísticas.

Methodological tools for (socio)linguistic analysis

Abstract
One of the most relevant steps in any scientific research is the choice of the methodological basis that will enable its realization. This choice depends, of course, on the object of study and on the theoretical approach adopted. Many theoretical models in linguistic studies invest in empirical analysis, facing the challenges generated by the manipulation of large volumes of data. Concerning (socio)linguistic studies, the methodology of the *Theory of Linguistic Variation and Change* stands out for the centrality attributed to obtaining, analyzing and interpreting data. In this paper, we present the computational methodological tools AntConc, Excel and Goldvarb X, which help in accomplishing all the fundamental steps for the study of linguistic variation and change.
Keywords: methodological tools; variation and change.

Introdução

Diante de vários caminhos viáveis para o desenvolvimento de investigações linguísticas, não é novidade a extrema importância de se tomar decisões adequadas quanto ao modelo metodológico a ser adotado e, feita essa escolha, de buscar o melhor caminho e os melhores recursos para executá-lo. Essa tarefa se mostra ainda mais desafiadora quando a pesquisa lida com grandes volumes de dados, característica inerente aos estudos variacionistas. Esses estudos envolvem a coleta, a organização e a análise/interpretação

de dados reais em amostras representativas da língua em uso, seja numa perspectiva sincrônica ou diacrônica.

Neste artigo, propomos apresentar ferramentas computacionais que permitem tornar esses estágios da pesquisa mais ágeis e precisos, possibilitando que o investigador disponha de mais tempo para se dedicar ao entendimento fidedigno da heterogeneidade constitutiva das línguas humanas.

Para tal, este texto está organizado em três seções. A primeira trata brevemente dos pressupostos teórico-metodológicos que fundamentam a Teoria da Variação e Mudança Linguísticas (WEINREICH; LABOV; HERZOG, 2006[1968]; LABOV, 1982, 1994, 2001, 2003, 2006[1966], 2008[1972], 2010), com destaque para a visão sociointeracionista da língua e o porquê dessa teoria priorizar um modelo de análise que opera com a quantificação de dados. Na sequência, apresentamos três ferramentas metodológicas – AntConc, Excel e Goldvarb X –, relacionadas, nesta devida ordem, às etapas de extração, organização e análise quantitativa. Por último, tecemos considerações finais, ressaltando positivamente a preocupação do pesquisador em investir na busca de novos conhecimentos para qualificar seus estudos.

Pressupostos teórico-metodológicos

No campo dos estudos da linguagem, talvez mais do que em outros domínios de estudo, o objeto é dado a partir da posição teórica adotada. Assim, cada modelo teórico pressupõe uma certa concepção de língua. Dentre as várias existentes, a Teoria da Variação e da Mudança Linguísticas concebe o seu objeto de estudo, a língua falada ou escrita, em seu contexto social real, tratando a variabilidade como algo intrínseco à linguagem humana e a caracterizando como regular e sistemática. Naturalmente, diferentes concepções teóricas têm implicações diretas nas escolhas metodológicas. Em seguida, exploramos essas duas ideias: a língua como entidade heterogênea e a necessidade de análises quantitativas de dados para apreender a sistematicidade subjacente a essa heterogeneidade.

Visão sociointeracionista da língua

A visão sociointeracionista da língua a concebe como um fenômeno "encorpado". Por "encorpado" entendemos, seguindo Marcuschi (2008), a articulação de aspectos sistemáticos (forma) e o funcionamento social, cognitivo e histórico que definem a língua.

Dentro dessa visão, a noção de heterogeneidade abrange três níveis distintos e interdependentes: a heterogeneidade do sistema linguístico, a heterogeneidade na comunidade linguística e a heterogeneidade de estilos.

Em relação ao primeiro nível de heterogeneidade, o modelo variacionista se distingue de outras abordagens por entender que o sistema linguístico não opera apenas com regras categóricas ou regras opcionais, mas inclui regras variáveis motivadas tanto por fatores linguísticos como extralinguísticos (LABOV, 2003, 2008[1972]).

No nível da comunidade, a heterogeneidade inerente ao sistema se materializa refletindo aspectos sócio-históricos e culturais dos grupos (distinções de classe social, papéis sociais de gênero, idade, etnia e origem geográfica). É pelo fato de a comunidade ser um espaço social heterogêneo que a variação também acontece em uma dimensão

entre falantes,[1] enquanto elementos representativos/constitutivos desses grupos. Essa dimensão também pode ser denominada "interfalante". Formas linguísticas variantes típicas de certos grupos podem ser convencionalmente associadas a esses segmentos como marcas identitárias. Os valores (significados) atribuídos às formas são postos em evidência quando da escolha de uma ou outra variante pelos usuários da língua.

O terceiro nível, o da heterogeneidade de estilos, situa-se na instância do falante, que, a depender da situação comunicativa da qual participa, opta por diferentes variantes linguísticas. Nesse nível, deve ser mensurado o efeito das características situacionais que condicionam o contexto comunicativo: onde o falante se encontra, com quem fala, por que fala, sobre o que fala. A essa dimensão nomeia-se "intrafalante".

Uma das preocupações do modelo teórico aqui discutido é definir estratégias para analisar a diversidade linguística proveniente dessas três heterogeneidades, características de toda língua. Em função da natureza dessas diferenças, uma compreensão integral dos resultados necessariamente conjuga um enfoque qualitativo e quantitativo, tendo sido esse último um dos diferenciais da proposta metodológica variacionista.

Por que análises quantitativas?

A necessidade de analisar quantitativamente os fenômenos linguísticos variáveis emerge da constatação, como já mencionado, de que não é possível chegar a uma descrição plena da língua em termos categóricos. Para Chambers (1995), a variável é variante, contínua e quantitativa. Segundo ele,

> Ela é variante no sentido de que é realizada diferentemente em diferentes ocasiões. É contínua no sentido de que certas variantes, tais como as gradações vocálicas para (eh) [...], assumem significância social dependendo de sua distância fonética em relação à variante padrão, ou, como no caso das variantes de (r), do quão diferentes são foneticamente da variante padrão. É quantitativa no sentido de que a sua significância não é meramente determinada pela presença ou ausência de suas variantes, mas por sua frequência relativa. (CHAMBERS, 2003, p. 26, tradução nossa)[2].

Assim, para estudarmos a variação, torna-se imprescindível medir o quão diferentes e o quão iguais são os usos linguísticos dentro da comunidade (MENDES, 2010). Essa análise quantitativa permite aferir os três níveis de heterogeneidade – do sistema, na comunidade e de estilos –, discutidos na subseção anterior.

[1] Neste texto, usamos o termo "falante" para englobar todo e qualquer enunciador. Isto é, alguém que produza um texto, seja falado ou escrito.

[2] "It is variant in the sense that it is realized differently on different occasions. It is continuous in the sense that certain variants, such as the vowel gradations for (eh) […], take on social significance depending upon their phonetic distance from the standard variant, or, as with the variants for (r), their phonetic differentness from the standard variant. It is quantitative in the sense that its significance is not determined merely by the presence or absence of its variants but by their relative frequency."

Ferramentas metodológicas: otimizando a análise da variação

Nesta seção, apresentamos as três ferramentas computacionais – AntConc, Excel e Goldvarb X –, explorando em cada uma delas os recursos que podem auxiliar no cumprimento de todas as etapas fundamentais para a concretização de estudos da variação e mudança linguísticas.

Uso do AntConc na extração de dados variáveis

O AntConc é um concordanciador, um programa computacional de uso livre, criado por Laurence Anthony, linguista e professor na University of Waseda, Japão (http://www.laurenceanthony.net/software/antconc/). Um concordanciador permite listar as ocorrências de uma determinada palavra ou frase em uma quantidade definida de contextos. De forma geral, também executa outras funções, como listar palavras de um texto ou *corpus,* extrair palavras-chave e colocados.[3]

Os recursos que compõem o AntConc são: *Concordance*; *Concordance Plot*; *File View*; *Clusters/N-Grams*; *Collocates*; *Word List* e *Keyword List*.

Conforme o Quadro 1, apresentamos um roteiro de utilização de alguns deles – *Word List, Concordance* e *File View* –, os quais simplificam e dinamizam a etapa de coleta de dados. O *Word List* mostra todas as palavras do *corpus* e as apresenta em uma lista ordenada. Isso permite que descubramos rapidamente quais palavras são as mais frequentes em um *corpus* e, em especial, se o fenômeno de interesse do pesquisador aparece nele. O *Concordance* apresenta o termo pesquisado e as linhas de concordância, ou seja, os fragmentos textuais em que o termo ocorre. O *File View* evidencia o termo pesquisado em um contexto estendido, já que ele aparece no texto específico que o comporta.

Quadro 1. Passo a passo para a utilização do AntConc

I)	PREPARAÇÃO PARA CARREGAR UM OU MAIS TEXTOS NO PROGRAMA:
1.	Prepare um arquivo de dados (= texto(s) que será(ão) analisado(s)) em formato *Word*.
2.	Salve-o em formato .txt (*Texto sem Formatação*) e na codificação de texto *Unicode (UTF-8)*.
II)	SELECIONANDO O(S) TEXTO(S) PARA A ATIVIDADE DE EXTRAÇÃO DE DADOS:
1.	Clique em *Open File(s)...* e selecione um ou vários arquivos em seu computador. - Para selecionar vários arquivos, mantenha a tecla "Ctrl" pressionada e, então, clique nos arquivos que quiser.
OU	
2.	Clique em *Open Dir...* para abrir todos os arquivos .txt de um diretório (pasta).
3.	Verifique o número de arquivos selecionados.
III)	INICIANDO A ATIVIDADE DE EXTRAÇÃO DE DADOS:
1.	Clique na aba *Word List*.
2.	Clique em *Start* e a lista de palavras em ordem de frequência será gerada.
3.	Para saber se determinada palavra está na lista, escreva-a no espaço indicado e clique em *Search Only*.
4.	Existem várias possibilidades de organização da lista (*Sort by...*).

(continua)

[3] Para maiores informações, cf. Sardinha (2004).

IV) REALIZANDO A EXTRAÇÃO DE DADOS:

1. Clique na palavra que deseja pesquisar. A aba *Concordance* abrirá automaticamente.
2. Para definir o tamanho das linhas de concordância, selecione os múltiplos de 5, que equivalem ao total de caracteres, inclusive espaços, igualmente distribuídos tanto à esquerda quanto à direita.
3. Todos os registros da palavra pesquisada aparecerão em azul e centralizados. Ao clicar em qualquer um deles, a aba *File View* abrirá. Nesse momento, aparecerá o contexto maior no qual a palavra desejada está inserida.
4. Se o dado visualizado for válido para os propósitos da pesquisa, salve-o ("Ctrl+C" / "Ctrl+V") em um outro arquivo (formato *Word* ou *Excel*).

Apresentamos, na sequência, as figuras 1 a 3 para ilustrar as etapas de extração de dados, envolvendo, respectivamente, os recursos *Word List, Concordance* e *File View.*

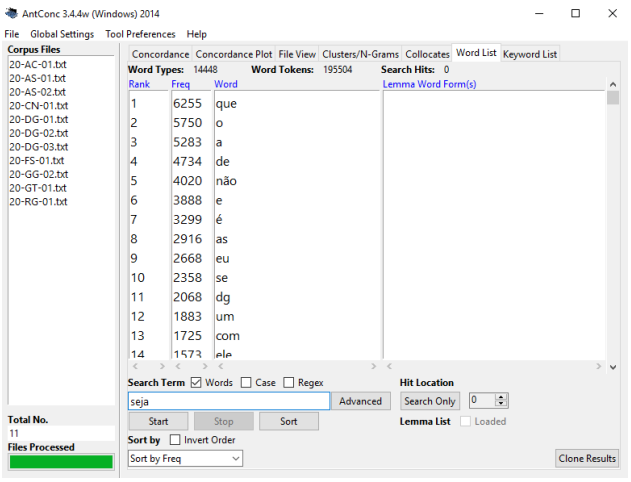

Figura 1. Extração de dados: recurso *Word List* em foco

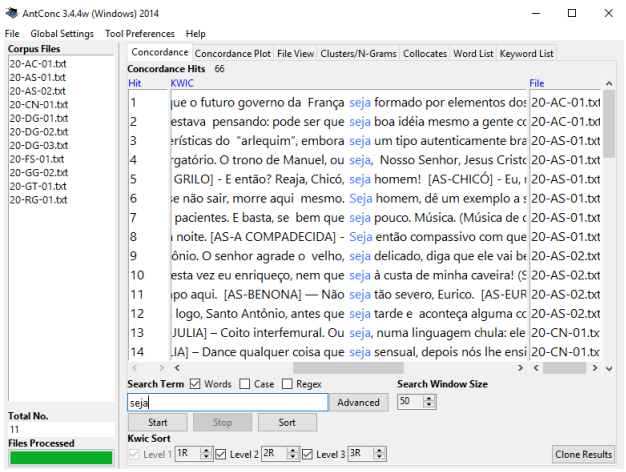

Figura 2. Extração de dados: recurso *Concordance* em foco

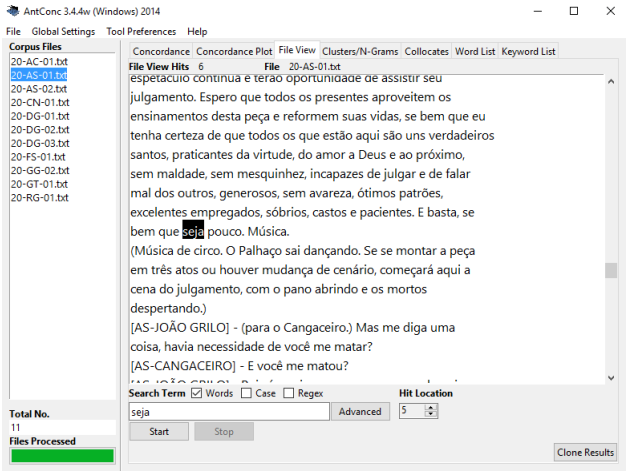

Figura 3. Extração de dados: recurso *File View* em foco

Uso do Excel na organização de dados variáveis

Excel ou Microsoft Excel é um aplicativo de criação de planilhas eletrônicas. Foi criado, em 1987, pela Microsoft. O Excel é muito prático para que sejam feitos controles, cálculos, tabelas; para que sejam colocados números em ordem crescente, uma lista de nomes em ordem alfabética, entre outras funções.

No âmbito das análises sociolinguísticas, o aplicativo é extremamente funcional para a preparação de arquivo de dados, codificação dos grupos de fatores,[4] geração de arquivo para ser utilizado no programa Goldvarb X. Além disso, o aplicativo oferece vários recursos para correção e localização de dados e criação de subconjuntos para análises específicas.

No Quadro 2, detalhamos as duas primeiras etapas tal como aplicadas pelo programa.

Quadro 2. Passo a passo para a utilização do Excel: parte 1

I)	PREPARAÇÃO DE ARQUIVO DE DADOS:
1.	Prepare um arquivo de dados em formato Word. Atenção à formatação: não deixe espaçamento de linhas entre os dados, nem dê "enter" entre os dados.
2.	Abra uma pasta Excel.
3.	Selecione os dados desse arquivo. Copie.
4.	Cole os dados na pasta Excel. Basta selecionar uma célula e dar o comando de "colar". O programa vai distribuir os dados na coluna, inserindo cada dado em uma linha.
5.	Recomenda-se inserir os dados na segunda linha de uma coluna (A ou B). A primeira linha pode ser usada para fazer um cabeçalho.
II)	CODIFICAÇÃO DOS GRUPOS DE FATORES:
1.	Para a codificação dos grupos de fatores, insira colunas antes da coluna de dados, uma para cada grupo.

(continua)

[4] O termo "grupos de fatores" se refere aos parâmetros selecionados como hipóteses explicativas para o fenômeno variável em estudo. Também se usa a expressão "variável independente" para designar esses parâmetros.

2.	Reserve a coluna A para o “parêntese” que abre o *code string*. Digite um parêntese dentro da célula A2. Para reproduzir o símbolo nas demais linhas dessa coluna, basta selecionar a célula e “puxar” a informação para baixo, até a última linha que contém um dado. Esse é um procedimento muito útil para a codificação.
3.	A coluna B vai conter sua variável dependente.
4.	Recomenda-se que você codifique cada grupo de fatores de uma vez, integralmente para o conjunto de dados. Isso garante maior consistência na análise.
5.	Você pode inserir novos grupos de fatores facilmente, se necessário, usando o recurso de “inserir coluna” do *Excel*, tal como assinalado na Figura 4.

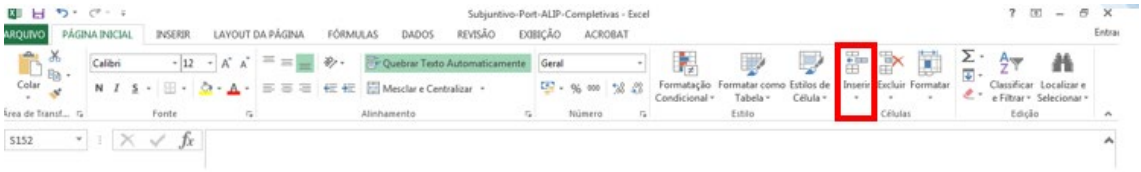

Figura 4. Recurso “inserir” do Excel

No Quadro 3, apresentamos a sequência de passos a serem executados para gerar um arquivo de dados codificados que será a base para os cálculos feitos pelo programa Goldvarb X. A Figura 5 ilustra a interface que se gera pelo uso da função específica para esse fim.

Quadro 3. Passo a passo para a utilização do Excel: parte 2

III)	GERAÇÃO DE ARQUIVO PARA *GOLDVARB X*:
1.	Uma vez completada a tarefa de codificação, usa-se a função “CONCATENAR”, para criar o arquivo com *code strings* necessário para a quantificação por meio do *Goldvarb X*.
a)	Selecione a primeira linha da coluna que vai conter os dados concatenados (geralmente a 2ª linha da coluna que segue imediatamente a coluna com os dados);
b)	Clique no ícone “função” na página inicial ou faça o seguinte caminho: **Fórmula > Inserir Função**.
2.	Acionada a função “CONCATENAR”, deve-se digitar a sequência de células (coluna + linha) que vão compor a sequência *code string* + dados: A2; B2; C2;.....X2. Se houver alguma coluna vazia ou que ainda não está totalmente completa ou que não se deseja incluir no cálculo a ser feito, digitar “” (duas aspas duplas) – ver a representação na figura 5. Antes de incluir a última célula – aquela que contém os dados, é necessário incluir um espaço entre os códigos e o dado em si. Para isso, digite aspas duplas, espaço e aspas duplas (“ ”).
3.	Com a sequência digitada, clique OK e a primeira sequência *code string* + dado estará pronta. Para gerar todas as sequências para o conjunto de dados, selecione a célula com a primeira sequência e “puxe” para baixo, até chegar ao último dado. Quando soltar a seleção, o programa terá aplicado a fórmula para todos os dados.
4.	Para levar esse conjunto para o *Goldvarb X*, selecione o conteúdo da coluna, copie e cole no *Goldvarb X*.

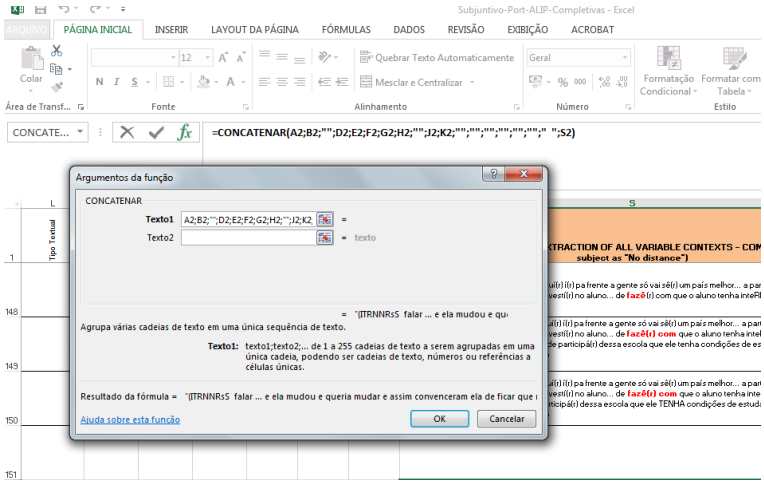

Figura 5. Uso da função *concatenar* do Excel

Como mencionamos antes, o Excel é um aplicativo muito versátil na gestão dos nossos dados. Dois recursos são particularmente úteis para as tarefas de correção e localização de dados e para a criação de subconjuntos: o uso de filtros e o recurso de ocultar/exibir colunas. Por exemplo, durante a análise da variação entre formas verbais de subjuntivo e indicativo, é possível que o pesquisador queira avaliar, do conjunto total de dados, apenas aqueles que reúnem algumas características, como terem por regente o verbo *acreditar* e o tempo pretérito na oração encaixada (*Eu acreditava que ele viesse*). Para obter esses dados específicos, podemos utilizar o recurso do filtro e, para visualizar melhor a operação, as colunas referentes a grupos de fatores não considerados no momento em questão podem ser ocultadas. No Quadro 4, fornecemos as instruções para a utilização desses dois recursos.

Quadro 4. Passo a passo para a utilização do Excel: parte 3

IV)	RECURSOS PARA CORREÇÃO, LOCALIZAÇÃO DE DADOS, CRIAÇÃO DE SUBCONJUNTOS – FILTROS:
1.	Para ativar os filtros nas colunas que contêm as codificações dos grupos de fatores:
a)	Selecione uma célula da primeira linha (em que temos o cabeçalho dos grupos);
b)	No menu da página inicial, selecione **Classificar e Filtrar > filtro**.
c)	Com isso é possível verificar se há células vazias (dados não analisados), quais são os códigos (categorias) presentes em cada grupo (como ilustrado na figura 6) e selecionar subgrupos com algumas características (combinando a seleção de códigos de grupos diferentes).
2.	Para o recurso de ocultar/exibir: selecionar a(s) coluna(s) que se quer ocultar ou reexibir, clicar com o botão direito do *mouse* e marcar "ocultar" ou "reexibir". É um ótimo recurso para "focalizar" um grupo de fatores que está sendo trabalhado, por exemplo.

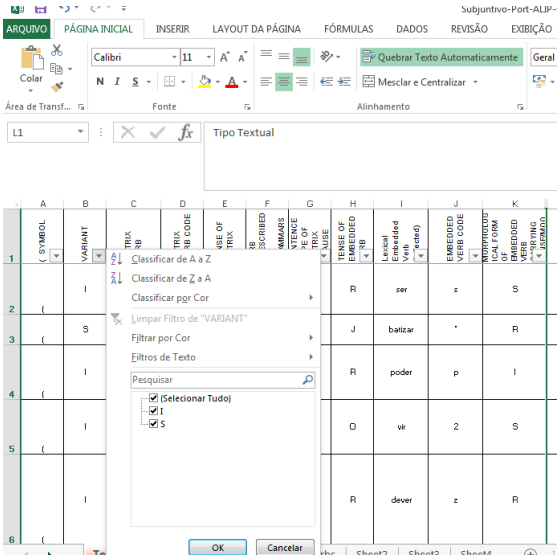

Figura 6. Uso da função *filtro* do Excel

Uso do Goldvarb X na quantificação/análise de dados variáveis

O Goldvarb X, programa estatístico computacional desenvolvido por Sankoff, Tagliamonte e Smith (2005) (http://individual.utoronto.ca/tagliamonte/goldvarb.html), é uma das ferramentas-chave da Sociolinguística Variacionista, em termos metodológicos.[5] Cabe a esse programa processar um grande volume de dados linguísticos, com o objetivo de definir uma regra variável que ajude a explicar determinado fenômeno sociolinguístico.

A partir de seu uso, podem ser feitas análises univariadas (ou unidimensionais), análises multivariadas (ou multidimensionais) e tabulações cruzadas. As análises univariadas são casos em que se testam o efeito de uma variável independente sobre uma variável dependente. Tais resultados aparecem sob a forma de frequências absolutas e relativas. As multivariadas permitem investigar situações em que a variável linguística em estudo é influenciada por vários elementos do contexto, ou seja, múltiplas variáveis independentes. Essa investigação mede os efeitos, bem como a significância dos efeitos, dessas variáveis independentes sobre a ocorrência das realizações da variável dependente. Os resultados obtidos se apresentam como pesos relativos. A tabulação cruzada, por sua vez, mostra as relações – ou a falta delas – entre as variáveis independentes (GUY; ZILLES, 2007).

No Quadro 5, disponibilizamos um roteiro simplificado para o uso do Goldvarb X, seguindo a ordem das etapas de análise cumpridas pelo programa: (i) geração do arquivo de dados; (ii) checagem da codificação; (iii) geração do arquivo de condições; (iv) análise univariada; (v) análise multivariada; e (vi) tabulação cruzada. Nesse processo, o programa vai gerar vários arquivos, cujas janelas devem permanecer sempre abertas enquanto o aplicativo estiver sendo executado.

[5] Temos que considerar, atualmente, a expansão no uso do Programa R (R TEAM, 2017) em estudos sociolinguísticos, o qual retomaremos nas conclusões.

Quadro 5. Passo a passo para a utilização do Goldvarb X: parte 1

I) GERAÇÃO DO ARQUIVO DE DADOS:
- Nesse arquivo, registram-se, após um parêntese, os códigos dos dados.
- Deve haver espaço considerável entre a sequência de códigos e o registro da ocorrência (esse espaço pode ser posto quando os dados forem organizados no *Excel*, como ficou especificado no quadro 3, item 2).
- Usualmente, a variável dependente costuma vir na primeira posição do *code string*.
- Para transportar um arquivo de dados, selecione o conteúdo da coluna referente à codificação pronta (no *Excel*) >> Copie >> Abra o programa *Goldvarb X >> Edit >> Paste*.
- Verifique se, ao transportar os dados, todos os parênteses que iniciam o *code string* estão alinhados na 1ª posição e se não há algum parêntese que não indica *code string* nessa posição. O programa quantifica todas as informações que seguem parênteses em 1ª posição.
- Salve o arquivo com a extensão .tkn, dando-lhe um título que expresse de modo significativo o fenômeno em análise.

II) CHECAGEM DA CODIFICAÇÃO:
- Essa etapa tem por objetivo verificar se não houve algum erro na hora da codificação. No menu selecione *Tokens >> Generate Factor Specifications*.
- Ao dar o comando *Generate Factor Specifications*, o programa preenche o quadro *Factor Specification*. Ali estão registrados: o número de grupos de fatores (ou variáveis independentes) e a relação de códigos definidos para os fatores de cada grupo. Deve-se checar se a lista de códigos de cada grupo corresponde fielmente ao conjunto definido pelo pesquisador.
- No caso de constatar códigos que não foram considerados (causados por erro de digitação, por exemplo), selecione *Tokens >> Find and Replace*. Em *Find and Replace*, digite o código que pretende achar e marque 'wrap around' e a coluna em que ele se localiza (a que grupo de fatores pertence) (cf. figura 7). A localização do código a ser corrigido vai ser indicada na tela do arquivo de dados. Depois de corrigir, salve o arquivo. Refaça a operação *Generate Factor Specifications* para verificar se ainda há erros.
- A última etapa de checagem é feita selecionando *Tokens >> Check Tokens*. Se todos os grupos tiverem pelo menos dois fatores, você lerá uma mensagem como "checking of tokens completed. XXX tokens in XXXX lines".
- Para visualizar, salvar e imprimir o arquivo de especificações: *Tokens >> Show Factor Specifications*. Essa será a primeira informação salva com a extensão .res.

III) GERAÇÃO DO ARQUIVO DE CONDIÇÕES:
- É o arquivo em que o pesquisador informa ao programa como quer que o arquivo de dados seja configurado para as análises.
- Esse arquivo pode ser criado **automaticamente** ou manualmente. Se escolhida a primeira opção, selecione *Tokens >> No Recode*. Salve o arquivo criado com a extensão .cnd.

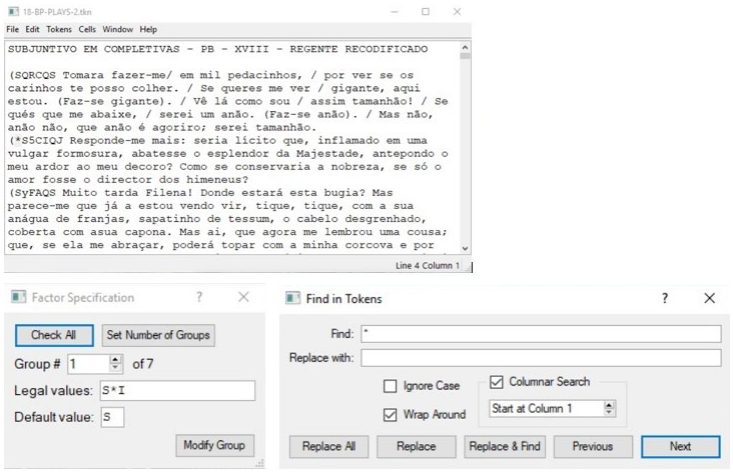

Figura 7. Uso da função *Find and Replace* do Goldvarb X

Quadro 6. Passo a passo para a utilização do Goldvarb X: parte 2

IV) ANÁLISE UNIVARIADA:

- Para realizar a análise univariada, clique em *Cells >> Load Cells to Memory*. Por meio desse comando, as condições contidas no arquivo de condições são aplicadas aos dados, gerando um arquivo de células, com extensão .cel, e um arquivo de resultados, com extensão .res.
- O arquivo de células (.cel) é o *input* ao programa de regra variável. As informações apresentadas são feitas para que o computador possa lê-las (e não os usuários).
- Antes de gerar o arquivo de resultados (.res), o pesquisador deve definir qual será o valor de aplicação (cf. figura 8). O valor de aplicação da regra é o primeiro código indicado na janela. Retomando o exemplo da variação entre as formas de subjuntivo e indicativo, se o foco da pesquisa é saber o quanto o indicativo já avançou em contextos tradicionais de subjuntivo, o valor de aplicação deve ser o código correspondente ao indicativo. Essa escolha condicionará o modo como serão apresentados os resultados das análises multivariadas.
- No arquivo de resultados (.res), cada grupo de fatores é identificado por dois números, por exemplo (cf. figura 9):

1 (2), "1": indicativo da variável independente no arquivo de condições;
"2": indicativo da posição da mesma variável no arquivo de dados.

- Quando fatores apresentam aplicação categórica (100% ou 0%), isso vem indicado como *knockout* (cf. figura 9). Como a análise multivariada só leva em conta casos de variação, esses dados categóricos precisam ser excluídos ou amalgamados a outros fatores. A escolha de uma ou outra opção dependerá da análise do pesquisador.
- Para eliminar *knockouts*, clique *Tokens >> Recode Setup*. Observar cada grupo, copiar quando estiver certo. No grupo em que ocorrer o *knockout*, excluir o fator "problemático" usando o comando *Exclude* no quadro da esquerda. Depois, copiar o grupo todo. Se a opção for amalgamar fatores, é preciso usar o recurso *Recode*.
- Rodar novamente: *Cells >> Load Cells to Memory*.

V) ANÁLISE MULTIVARIADA:

- Para a realização da análise multivariada selecione *Cells >> Binomial Up and Down*.
- O programa fornece um resumo dos grupos selecionados como mais significativos para a realização de determinado fenômeno e, também, dos grupos excluídos.
- A melhor e a pior rodada são destacadas. (*Best stepping up run* e *Best stepping down run*).
- Para a interpretação dos resultados, os pesos relativos devem ser retirados da melhor rodada indicada pelo programa (cf. figura 10). No intervalo entre 0 e 1, pesos relativos de 0.5 são considerados neutros em relação à aplicação da regra variável. Valores abaixo de 0.5 indicam que o fator desfavorece a aplicação da regra; valores acima de 0.5 indicam que o fator favorece essa aplicação.

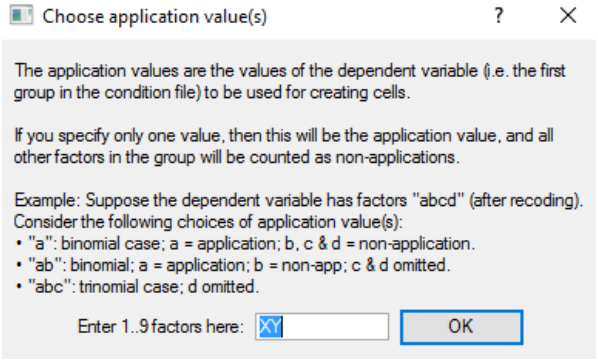

Figura 8. Escolha do valor de aplicação

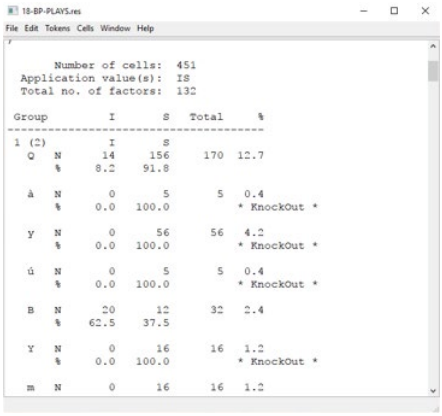

Figura 9. Uso da função *Find and Replace* do Goldvarb X

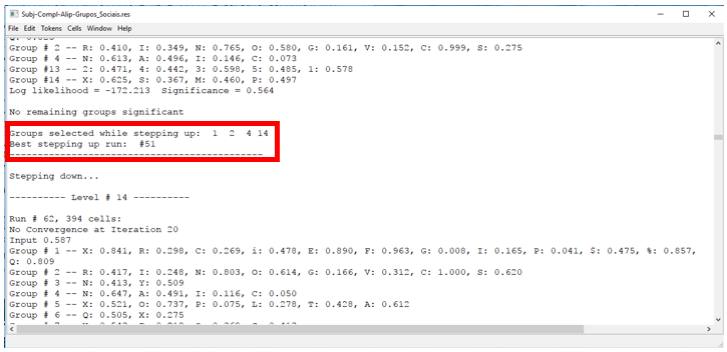

Figura 10. Resultado de *Stepping up* e indicação da melhor rodada

Quadro 7. Passo a passo para a utilização do Goldvarb X: parte 3

VI)	TABULAÇÃO CRUZADA: - Para cruzar os dados de dois grupos de fatores, combinando sua ação na escolha das variantes, selecione *Cells >> Cross Tabulation*. Aparece uma janela para que sejam digitados os grupos que devem ser cruzados. Deve-se tomar como referência os números que aparecem fora dos parênteses junto de cada grupo de fatores no arquivo de resultados da análise univariada (como indicado na figura 11).

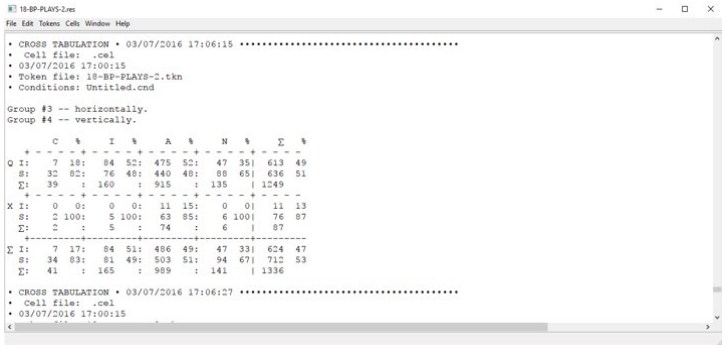

Figura 11. Resultado de Tabulação Cruzada

Considerações finais

Neste artigo, propusemos a utilização de um conjunto de três ferramentas computacionais que, quando articuladas, podem otimizar as tarefas de coleta, organização e quantificação de dados linguísticos. Cabe ressaltar que, se demos um destaque à utilização desses recursos em análises de fenômenos variáveis, as ferramentas podem ser bastante úteis também para o pesquisador cujo fenômeno em estudo não tem natureza variável, mas que implica, de todo modo, uma investigação de grandes volumes de dados.

A amplitude de uso das ferramentas nos parece evidente no que se refere ao aplicativo AntConc e, mais ainda, ao Excel. É fato que o programa Goldvarb X foi concebido especificamente para o estudo de fenômenos linguísticos variáveis, e que a aplicação total de seus recursos visa levar à identificação de propriedades e características que motivam a escolha de uma ou outra forma variante, em contextos em que tais formas são concorrentes no uso. No entanto, também no caso desse aplicativo não se pode deixar de reconhecer seu potencial para auxiliar na caracterização da distribuição de uso de formas linguísticas não variáveis. A partir da codificação de dados com base em parâmetros definidos como relevantes para sua caracterização, torna-se bastante simples por meio do Goldvarb mapear a frequência absoluta e relativa de cada forma, e suas interrelações com as categorias controladas. Trata-se de um uso heurístico amplo, que nos permite identificar padrões de uso.

Finalizamos essa discussão destacando que a área começa a explorar outras ferramentas, para além daquelas aqui mencionadas, como o programa R (https://cran.r-project.org/). Essa ferramenta já tem sido bastante utilizada em outros campos de pesquisa, por exemplo, nas ciências exatas. Na Linguística, parecemos vivenciar um período transitório, em que estamos testando o potencial desses instrumentos, com o intuito de avaliar em que medida são semelhantes, possuem recursos específicos e mais vantajosos, ou podem ser utilizados complementarmente.

REFERÊNCIAS

CHAMBERS, J. K. *Sociolinguistics Theory*: Linguistic Variation and its Social Significance. 2. ed. Oxford: Blackwell Publishers, 2003.

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

LABOV, W. *Principles of Linguistic Change*. v. 3: Cognitive and Cultural Factors. Oxford: Wiley-Blackwell, 2010.

_____. *Padrões sociolinguísticos.* São Paulo: Parábola, 2008[1972].

_____. *The Social Stratification of English in New York City*. Cambridge: Cambridge University Press, 2006[1966].

_____. Some sociolinguistic principles. In: PAULSTON, C. B.; TUCKER, G. R. (Eds.). *Sociolinguistics*: the essential readings. Oxford: Blackwell, 2003. p. 234-250.

_____. *Principles of Linguistic Change*. v. 2: Social factors. Cambridge: Blackwell, 2001.

_____. *Principles of Linguistic Change*. v. 1: Internal factors. Cambridge: Blackwell, 1994.

_____. Building on Empirical Foundations. In: LEHMANN, W. P.; MALKIEL, Y. (Eds.). *Perspectives on Historical Linguistics.* Philadelphia: John Benjamins Publishing, 1982. p. 17-92.

MARCUSCHI, L. A. *Produção textual, análise de gêneros e compreensão*. São Paulo: Parábola Editorial, 2008.

MENDES, R. B. A Variação Linguística. In: FIORIN, J. L. (Org.). *Introdução à Linguística I.* Objetos Teóricos. 6. ed. São Paulo: Contexto, 2010. p. 121-140.

R TEAM, Development Core. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing, 2017. Disponível em: <http://www.R-project.org/>. Acesso em: 09 fev. 2015.

SARDINHA, T. B. *Linguística de Corpus*. Barueri: Editora Manole, 2004.

SANKOFF, D.; TAGLIAMONTE, S. A.; SMITH, E. *Goldvarb X*: A variable rule application for Macintosh and Windows. Department of Linguistics, University of Toronto, 2005. Disponível em: <http://individual.utoronto.ca/tagliamonte/goldvarb.html>. Acesso em: 09 fev. 2015.

WEINREICH, V.; LABOV, W.; HERZOG, M. *Fundamentos empíricos para uma teoria da mudança linguística.* São Paulo: Parábola, 2006[1968].

Recebido em: 10/10/2017

Aprovado em: 21/03/2018