Léxico Rural da Serra do Cipó

Vitor de Castro Silva
Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Londrina (UEL)
Londrina, Brazil
vitor.castro.silva@uel.br

Cinthyan Renata Sachs Camerlengo de Barbosa
Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Londrina (UEL)
Londrina, Brazil
cinthyan@uel.br

Wagner Ferreira Lima
Departamento de Letras Vernáculas e Clássicas
Universidade Estadual de Londrina (UEL)
Londrina, Brazil
wflima@uel.br

ABSTRACT
As the expression of the identity of a people, the lexicon should be studied by itself. Today, with the growing wave of valuing the differences, the study of lexicons has become even more relevant. Interfaces must take this into account if they intend to respond to social demands. Therefore, a survey was conducted to promote and recognize the rural linguistic variety in interface research. The paper consisted, of indexing the rural vocabulary of the Serra do Cipó, metropolitan region of Belo Horizonte-MG, with hash function from Aho obtaining optimizations for all operations. The result was an online dictionary that collects information about the rural linguistic variety of the region. Finally, the conclusion is that such a dictionary represents an important means of digital inclusion and should serve as a model for further work.

Keywords
Digital inclusion; Serra do Cipó; Natural Language Processing.

ACM Classification Keywords
Human-centered computing: Human Computer Interaction (HCI).

RESUMO
Como expressão de identidade de um povo, o léxico deve ser estudado por si só. Atualmente, com a onda crescente de valorização das diferenças, o estudo dos léxicos se tornou ainda mais relevante. As interfaces devem levar isso em conta se quiserem responder às demandas sociais. Em vista disso, foi realizada uma pesquisa para promover o reconhecimento de variedade linguística rurais. O trabalho consistiu em indexar o vocabulário rural da Serra do Cipó, região metropolitana de Belo Horizonte-MG, com uma função hash do Aho obtendo otimizações para todas as operações. O resultado foi um dicionário *online* que reúne informações sobre a variedade linguística rural da região. Por fim, conclui-se que um dicionário como esse representa um importante meio de inclusão digital, devendo servir como um modelo para trabalhos futuros.

Palavras-chave
Inclusão Digital; Serra do Cipó; Processamento de Linguagem Natural.

INTRODUÇÃO

O desenvolvimento acelerado das tecnologias da comunicação e informação representa também um desafio a ser enfrentado pela sociedade brasileira, a despeito de seu grande avanço tecnológico. A mais evidente delas talvez seja a sensação generalizada de homogeneização e padronização culturais, ressaltada por Manzano [1].

O conceito de diversidade é central nessas discussões. Essa pode ser entendida como condição necessária à prática interdisciplinar e transdisciplinar representada pela diferença ou o não reconhecimento do outro, como igual a nós [2], em termos das ideias, crenças, costumes, etnias, classes sociais, linguagens, profissões, personalidades, etc.

Assim sendo, o objetivo deste trabalho é oferecer um modelo simples de implementação computacional de dados sociolinguísticos que possa permitir a inclusão da cultura de comunidades regionais em aplicações de Processamento de Linguagem Natural (PLN).

No caso, o aspecto cultural escolhido é um vocabulário rural estudado e divulgado por Freitas [18]: o léxico-cultural da Serra do Cipó/MG. Assim, os dados usados em nossa implementação são os encontrados originalmente nesse estudo.

Finalmente, usamos a conhecida função hash Aho [3] como método de indexação, devido à sua eficiência [4] e busca dos verbetes do referido léxico.

Este trabalho se justifica por duas exigências sociais atuais [5]: uma sociolinguística e outra de inclusão digital.

Quanto à primeira, o léxico é o aspecto linguístico que mais imediatamente sinaliza a visão de mundo de um povo. Isso porque a cultura de uma comunidade está registrada nas palavras empregadas por seus membros. É assim que, como veremos, o léxico dos moradores da Serra do Cipó traz significados marcantes sobre a cultural local.

Quanto à segunda exigência, a da inclusão digital, a presente apresentação pretende ser uma contribuição com as pesquisas em Interfaces em Linguagem Natural para Banco de Dados (ILNBDs). Os trabalhos em interfaces se quiserem estar sintonizados com as exigências atuais da

sociedade, precisam integrar a cultura de seus usuários em seus modelos de interação homem-computador.

Este artigo se organiza como segue. Na Seção 2 são referenciados alguns trabalhos correlatos. Na Seção 3 é apresentado o trabalho de onde foi extraído o vocabulário. Na Seção 4 é descrito o método de *hashing* empregado na indexação desse vocabulário. Na Seção 5 os materiais e métodos usados na implementação da tabela hash são apontados. Na Seção 6, os dados resultantes da implementação são apresentados. Na Seção 7, há uma breve discussão sobre esses resultados. Por fim, as considerações finais estão na Seção 8.

TRABALHOS CORRELATOS

Léxicos estruturados e o formalismo gramatical são centrais para um sistema de PLN. Muitos trabalhos sobre léxicos têm sido apresentados, porém nem todos nos dão informação semântica precisa. Fargetti, Murakawa e Nadim [6] chamam a atenção sobre a dificuldade do leitor quando em dicionários monolíngues (o mesmo acontece com os bilíngues), ao se deparar com remissivas que, longe de esclarecer o significado de um item lexical, em português, lança mais dúvidas por sua falta de detalhes.

Em sua concepção tradicional, o léxico é uma lista não-estruturada de "palavras" ou "entradas", contendo, para cada uma delas, a especificação de sua realização fonética, de suas propriedades morfológicas, sintáticas e semânticas, além de conter todas as idiossincrasias, isto é, formas que não podem ser previstas como o resultado da aplicação de princípios da morfossintaxe [7].

Do ponto de vista psicolinguístico, a produção de enunciados envolve três tipos de processos mentais: (1) especificação de conceitos (conceptualização, C), (2) seleção de palavras e construção de representações sintáticas e fonéticas (formulação, F) e (3) produção da fala (articulação, A). Por exemplo, o processo de nomeação de um objeto perceptível envolve a identificação do objeto (C), a seleção de uma representação sintático-semântica do objeto, bem como a codificação dessa representação em termos fonológicos (F), e, finalmente, a transformação dessa representação fonológica em realização fonética, que constitui o nome do objeto [7].

Porém, ainda temos a Sociolinguística [8] que é um ramo da Linguística que se caracteriza por conceber uma língua como uma realidade essencialmente sociocultural. Em seu cerne está a hipótese de que as variações observáveis na linguagem verbal são processos regulares e *ipso facto* passíveis de estudo e sistematização.

Muitos trabalhos sobre léxico têm sido publicados, porém são de conceitos mais específicos, como Léxico das Orquídeas [9], Léxico das Ervas [10] e Léxico das Pragas das Sojas [11].

Segundo Paim [12] e Souza [13], ferramentas e métodos de consultas para léxicos digitais são bem escassos para a Língua Portuguesa. Para Greghi [14] a criação de aplicações desse tipo é uma tarefa árdua e demorada que poderá ficar simplificada com a centralização dos dados em repositório que armazene todas as informações lexicais. Outro problema destacado é a maneira em que essas informações são organizadas e o tratamento dado a essas bases em suas construções que poderão acarretar em baixos desempenhos nas buscas.

Trabalhos sobre léxicos rurais têm sido estudados. Marins [15] discute vestígios de ruralidade no léxico dos habitantes da região Centro-Oeste do Brasil com base em dados geolinguísticos têm sido estudados e documentados pelo Projeto *Atlas Linguístico do Brasil* (ALiB). O antigo norte de Goiás, atualmente Tocantins, também foi abordado em Silva e Borges [16] pelos seus traços de ruralidade. Ambos trabalhos foram apoiados nas seguintes áreas: lexicologia, dialetologia e geolingüística. Essas áreas da linguística fornecem subsídios para o referido trabalho.

De modo similar, trataremos neste artigo o léxico-cultural da Serra do Cipó da região Sudeste.

LÉXICO-CULTURAL DA SERRA DO CIPÓ

Léxicos são bases de grande volume de dados que têm em seu conjunto vários atributos linguísticos (etimologia, pronúncia, morfologia, sintaxe e definições) para cada um dos seus itens lexicais que podem servir a uma aplicação ou apenas como centralização e organização das informações [17].

A relação necessária entre língua, cultura e sociedade que fazemos questão de enfatizar é relevante para os dias de hoje. De um lado, ela cumpre uma exigência das sociedades contemporâneas, de que as culturas regionais sejam respeitadas. De outro, ela sugere que a diversidade sociolinguística deve ser uma preocupação das ILNBDs [5].

A elaboração de uma interface como o léxico do café da Serra do Cipó está em linha com essas exigências atuais da sociedade. As informações para a indexação digital foram extraídas de Freitas [18] que realizou um justo estudo sociolinguístico sobre o vocabulário rural da região da Serra do Cipó, localizada na região Metropolitana de Minas Gerais. A área pesquisada abrange parte de dois municípios mineiros: Jaboticatubas e Santana do Riacho, os quais se caracterizam por seu caráter predominantemente rural.

Essa região foi escolhida por diversos fatores sociolinguísticos, entre os quais as peculiaridades percebidas na fala dos moradores. Trata-se de uma variedade linguística do português que simboliza a cultura rural local da região. Freitas [18] ressalta não existirem registros de realização de estudos de cunho lexical focados nessa região. Esse léxico é uma contribuição com os

estudos dialetológicos e lexicais sobre comunidades presentes em território mineiro.

A regularidade das variações é explicada em termos dos condicionamentos sociolinguísticos. De acordo com isso, fatores socioculturais como a geografia, faixa etária, classe social, etc. controlam o uso que os usuários fazem de formas linguísticas. Essa hipótese se aplica também ao léxico de uma língua.

Dentro da linguística, o léxico é visto como o componente no qual mais se notam as influências socioculturais sobre a língua. Os trabalhos americanos em Antropologia no início do século 20 foram incisivos em defender a inter-relação entre língua, sociedade e cultura. E desde então muitos trabalhos têm enfatizado esse inter-relacionamento [8].

Em geral, eles têm evidenciado que o vocabulário usado por uma comunidade modela o jeito pelo qual seus membros vão conceber e experimentar a realidade. O léxico rural da Serra do Cipó é uma maior evidência desse fenômeno.

Cumpre dizer que um estudo lexical pode ser lexicológico ou lexicográfico, a diferença sendo quanto aos métodos e fins assumidos. Em termos gerais, a lexicologia visa ao estudo da palavra no sentido de categorizá-la e de analisar sua estrutura dentro do universo lexical. Já a lexicografia se dedica à prática dicionarística, ou seja, à produção de dicionários, glossários e vocabulários [19].

O dicionário objetiva reunir e definir o maior número possível dos lexemas de uma língua. O vocabulário, por sua vez, procura dar conta do conjunto de lexemas de um determinado tipo de discurso (político, geográfico ou religioso); como é o caso dos vocabulários técnico-científicos e especializados. Finalmente, o glossário se caracteriza por ser um esclarecimento do contexto lexical de um único texto, ou obra, manifestado.

O trabalho de Freitas [18] descreve assim o vocabulário dos moradores da Serra do Cipó e estabelece um glossário acerca do emprego desse vocabulário. A pesquisa foi totalmente baseada em entrevistas orais de 12 moradores da região, as quais foram transcritas, conforme método sociolinguístico apropriado. Das transcrições foram selecionadas lexias que melhor representassem a realidade da população local [18].

Uma ficha lexicográfica foi elaborada para cada lexia selecionada, contendo informações relativas à sua definição e etimologia. Também foi elaborado um glossário com o intuito de sistematizar a consulta a tais vocábulos.

Como veremos, é esse glossário que foi mapeado em uma estrutura de dados hash, para que ele possa ser acessado por meio de algum dispositivo eletrônico conectado à Web.

A indexação desse glossário em uma estrutura de dados digital tem o efeito de permitir o reconhecimento da identidade cultural dos moradores da Serra do Cipó. Porém, ela pode ser vista também como a etapa inicial de um processo de interação homem-computador que envolva variedades linguísticas regionais, com objetivos práticos distintos.

Abordamos alguns dos trabalhos que sugerem ser perfeitamente possível construir interfaces em contextos específicos de atividade humana. No nosso caso, contudo, vamos ainda além; e apregoamos a construção de interfaces mesmo para variedades linguísticas regionais, como é o exemplo do léxico da Serra do Cipó.

Em suma, a indexação digital do léxico rural é relevante pelas diversas razões sociais e mesmo políticas, ora apontadas. Esse conhecimento é fundamental para a implementação de um banco de dados, mas é preciso também conhecer algumas questões técnicas que um projeto desse tipo coloca, a saber: qualquer aplicativo em ILNBD precisa ser eficiente em seu tempo computacional.

Atualmente a chamada função hash, ou também denominada de espalhamento, se mostra como um dos meios computacionais mais eficientes na indexação de dados. A seguir, vamos descrever uma técnica consagrada entre os cientistas da computação, conhecida como função Aho. Essa foi escolhida baseando-se nas implementações de 14 funções implementadas por Moreno [4] e discutidas por Moreno *et al.* [20], as quais são baseadas não só em pesquisas em Estrutura de Dados, mas também específicas para Linguagem Natural projetadas e sugeridas por Jenkins [21], que é um pesquisador da computação e autor de várias funções hash.

FUNÇÃO HASH

O acesso a um banco de dados extenso, como é o caso de um léxico, requer um algoritmo eficiente. E eficiência em ILNBD tem a ver basicamente com o tempo e precisão de processamento: acessar um léxico grande de modo rápido e otimizado [4]. Já há algum tempo a indexação por meio da função hash ou de espalhamento vem satisfazendo esse requerimento. A estrutura de dados decorrente do espalhamento é conhecida por *tabela hash*, ou tabela de dispersão.

Em linhas gerais, a função hash consiste em transformar dados a serem armazenados em índices, por meio dos quais esses dados podem posteriormente ser acessados. O diferencial da tabela hash está no modo como esse processamento tem lugar, espalhando as informações sobre um vetor de tamanho fixo. As principais características dessa função são:

(a) Ela toma qualquer dado de entrada, seja texto, inteiro, ponto flutuante, tupla, etc., e o converte em um valor numérico inteiro, no intervalo do vetor (m-1), onde os

dados serão armazenados. Em nossa demonstração, a função recebeu as tuplas (chave, valor) como entrada; as quais, uma vez convertidas, foram endereçadas a uma posição do vetor;

(b) Ela opera sobre um vetor que, por razões de performance, precisa ser maior que o número de chaves a serem indexadas: quanto mais espalhadas as chaves de entrada estiverem no intervalo do vetor, mais eficiente será a sua busca e inserção na tabela;

(c) A função computa os dados de entrada ou vocabulário a ser indexado, tendo em conta os espaços vazios existentes na tabela. A literatura recomenda calcular o chamado fator de carga a fim de estimar o equilíbrio dessa relação. O fator de carga é dado pela divisão da quantidade do vocabulário pelo tamanho da tabela. Foi escolhido o valor de 0.75, o qual é o mesmo utilizado por tabelas hash na linguagem Java.

Em suma, a eficiência atribuída a essa função se deve ao fato de as chaves serem distribuídas esparsamente pelo vetor; tal que quaisquer operações efetuadas por um algoritmo de espalhamento serão da ordem de O(1). Isso significa dizer que o algoritmo requer uma única comparação para encontrar a chave solicitada.

A aplicação do espalhamento apresenta alguns problemas. A colisão de dados é, senão o mais proeminente, pelo menos o principal deles. Colisão consiste no fato de duas ou mais chaves receberem o mesmo endereçamento no índice; isso em razão de os valores para essas chaves acabarem por algum motivo coincidindo-se. Essa é um fato inevitável que pode ser prejudicial ao processamento. Pode acontecer que a inserção de uma nova chave apague aquela que foi inserida anteriormente. Sendo assim, colisões devem ser necessariamente evitadas. Um corpo de pesquisas nessas últimas décadas tem trabalhado nessa direção, dando lugar a diferentes métodos de tratamento de colisões.

A seguir vamos descrever a função hash Aho que mostrou ser eficiente [10] tendo em conta sua otimização comparada com a das outras funções consideradas [4]. Por isso, optamos por ela nesta interface do léxico rural da Serra do Cipó.

A função Aho é a seguinte [3]:

1) Determinar um inteiro positivo *h* a partir dos caracteres c_1, c_2, ... c_k na cadeia s;

2) O valor antigo de "*h*" é então multiplicado por um α antes de o próximo caractere ser adicionado;

3) O valor de *hash* é o resto de *h mod m*, onde sugere-se que *m* seja um número primo.

Esse método de hash pode ser resumido na seguinte equação:

"$h = \alpha * h + (Chave[i])$"

Moreno [4] observa que usar valores de tabela de base 2 atrapalha a correta distribuição dos elementos e, também, no tempo dessa função. Assim sendo, foi escolhido um número primo para o tamanho da tabela. Esse valor primo também é recomendado por Aho em Moreno, Barbosa e Manfio [20].

Assim sendo, foi escolhido um número primo para o tamanho da tabela. Para o vocabulário da Serra do Cipó que contém 341 palavras diferentes, o tamanho estipulado foi 457, sendo esse o primeiro primo maior que 1,33 vezes o tamanho do vocabulário.

MATERIAIS E MÉTODOS

É possível separar a metodologia tomada em duas etapas. A primeira consistiu em pesquisar um léxico que simboliza a diversidade cultural em nível geográfico. O léxico da Serra do Cipó nos pareceu nesse sentido apropriado. Já a segunda etapa se constituiu efetivamente em escrever um programa para acessar a estrutura de dados usando a função hash.

Como vimos, os dados para a construção dessa estrutura estão disponíveis em Freitas [18]. Nesse estudo acadêmico, encontramos um glossário pronto, contendo os vocábulos usados pela comunidade rural da Serra do Cipó como meio de comunicação. Isso obviamente facilitou nosso trabalho, que basicamente consistiu em compilar os dados e indexá-los em uma tabela *online*.

Abaixo temos a definição das propriedades dos verbetes desse glossário, tal como encontrada em Freitas [18].

Na Tabela 1 a ordenação das propriedades dos verbetes, seguida da descrição da construção da estrutura.

LEXIA; dicionarização; categoria gramatical; origem; definição; abonação; número da entrevista e linha do corpus; (eventualmente) variação linguística. A lista de abreviaturas e convenções auxilia a leitura dos verbetes: "ABISAR • (A) •[V] • Lat>Port • Dar aviso a, informar, previnir. Variante de avisar. • *"ele foi abisado que tava com...mas ê num ligava pra coisa né pra pressão arta né."* (Ent.07,linha 180) • (abisar ~ avisar: caso de degeneração)" [18].

Abreviaturas e convenções	
A – dicionarizado no Aurélio	loc. pron – locução pronominal
adv. – advérbio	n/A – não-dicionarizado no Aurélio
afr. – africanismo	n/d – não-dicionarizado em nenhuma das obras consultadas
arc. – arcaísmo	n/e – não encontrada
Ár. – árabe	Mal. – malaia
Cast. – castelhana	Nap - napolitana
Cel - celta	NCf – nome composto feminino
Cf. - conferir	NCm – nome composto masculino
cont. – controvertida	Nf – nome feminino
desc. – desconhecida	Nm – nome masculino
duv - duvidosa	obs. – obscura
esp. – espanhola	onomat. – onomatopaica
Fr. - francesa	PESQ. – pesquisadora
Germ. – germânica	prep. – preposição
Greg. - grega	pron. – pronome
inc. – incerta	Ssing – Substantivo singular
ind. – indigenismo	top – toponímica
INF. – informante	V – verbo
lat. – latim	
loc. adv. – locução adverbial	

Tabela 1. Abreviações e convenções do glossário

Considerando-se a fórmula "*h = α*h + (Chave[i])*", que requer um *h* para cada iteração sobre a chave de caracteres, *h* vai ser assim atualizada em cada iteração. Esse cálculo é para reduzir os casos de colisões e, quando essas forem inelutáveis, também para evitar que muitas chaves ocupem o mesmo endereço.

Assim, uma classe foi criada, a *class TabelaHash*, a qual contém os métodos para gerar a função de espalhamento e, também, para operar sobre a tabela hash criada:

(a) Método para o espalhamento: *função_AHO( ):* Em conformidade com Moreno [4], adotamos α = 10. Essa função multiplica a hash antiga pelo alfa e depois soma o valor do caractere. Em seguida ela toma o resto da hash pelo primo mais próximo do tamanho da tabela (*m*). Finalmente, o resto desse valor pelo tamanho real da tabela é o hash da chave de entrada.

Para encontrar o primo, criamos as seguintes funções: *is_prime(number)* e *next_prime(number)*. A primeira determina se um número é primo; a segunda encontra o próximo primo maior que o número dado.

Dentro do método construtor da classe, o vetor foi determinado por uma função de compreensão de lista: *self.tabela_hash = [[ ] for i in range(self.tabela_size)]*. Já o valor de "tabela_size" foi obtido por meio do cálculo do fator de carga (tamanho_do_vocabulario / 0,75).

(b) Métodos para operações sobre a tabela: uma vez obtido o espalhamento, foi possível inserir, buscar e remover itens. A função *insert(self, chave: str, valor)* adiciona uma chave no vetor. Caso o item já exista, a função o retira da posição. A busca acontece mediante a função *get(self, chave)*, que recupera o item buscado. Por fim, *remove(self, chave)* remove do vetor o item considerado. Os resultados são apresentados a seguir.

Considerando-se a fórmula "*h = α*h + (Chave[i])*", que requer um *h* para cada iteração sobre a chave de caracteres, *h* vai ser assim atualizada em cada iteração. Esse cálculo é para reduzir os casos de colisões e, quando essas forem inelutáveis, também para evitar que muitas chaves ocupem o mesmo endereço.

Assim, uma classe foi criada, a *class TabelaHash*, a qual contém os métodos para gerar a função de espalhamento e, também, para operar sobre a tabela hash criada:

(a) Método para o espalhamento: *função_AHO( ):* Em conformidade com Moreno [4], adotamos α = 10. Essa função multiplica a hash antiga pelo alfa e depois soma o valor do caractere. Em seguida ela toma o resto da hash pelo primo mais próximo do tamanho da tabela (*m*). Finalmente, o resto desse valor pelo tamanho real da tabela é o hash da chave de entrada.

Para encontrar o primo, criamos as seguintes funções: *is_prime(number)* e *next_prime(number)*. A primeira determina se um número é primo; a segunda encontra o próximo primo maior que o número dado.

Dentro do método construtor da classe, o vetor foi determinado por uma função de compreensão de lista: *self.tabela_hash = [[ ] for i in range(self.tabela_size)]*. Já o valor de "tabela_size" foi obtido por meio do cálculo do fator de carga (tamanho_do_vocabulario / 0,75).

(b) Métodos para operações sobre a tabela:

Uma vez obtido o espalhamento, foi possível inserir, buscar e remover itens. A função *insert(self, chave: str, valor)* adiciona uma chave no vetor. Caso o item já exista, a função o retira da posição. A busca acontece mediante a função *get(self, chave)*, que recupera o item buscado. Por fim, *remove(self, chave)* remove do vetor o item considerado. Os resultados são apresentados a seguir.

RESULTADOS

Com uma quantidade dos verbetes indexados (430 expressões), foi possível ter uma noção da eficiência da implementação da referida tabela hash. Contudo, baseando-nos em pesquisas prévias (como trabalhos como Professor Tical de Manfio, Moreno e Barbosa [22] que também utilizou-se de tabelas hash de endereçamento encadeado obtendo ótimo desempenho), podemos afirmar que a função Aho possibilitou uma busca eficiente dos dados do glossário.

As três operações básicas (inserir, buscar e remover) sobre a tabela hash criada foram realizadas com sucesso (figuras 1, 2 e 3, respectivamente) permitindo assim que uma interface de baixo custo computacional com o usuário seja possível. Para comprovar a eficiência da tabela, também foram feitos testes de velocidade. A Figura 4 demonstra os resultados obtidos com testes de velocidade em microssegundos.

```
tabela.insert('ABISAR', {
    'palavra': 'Abisar',
    'tipo': 'V',
    'ex_uso': 'Tabaiava lá naquela época né...então e ele foi abisado
    'reg_bluteau': 'Avisar Fazer Aviso',
    'reg_morais': 'Avisár v.at. Dar, fazer aviso; noticiar; amoestar.'
    'reg_laud': 'Avisar v. r. v. B. lat. advisare. Dar aviso a, anunci
    'reg_aurelio': 'Avisar [Do fr. aviser] V.t.d. 1. Dar aviso a; faze
    'reg_cunha': 'Avisar vb. 'informar, prevenir'    avy- XIV   Do fr.
    'reg_amadeu': 'n/e',
    'glossarios': {'souza': 'n/e', 'ribeiro': 'n/e'}
```

Figura 1. Inserir informações com insert (self, chave: str, valor).

```
tabela.get('ABUSANTE')

{'ex_uso': 'Uai ô durmi na casa de (R...)...(R...L...)...(L...) do Morro fui buscá ma
 'glossarios': {'ribeiro': 'n/e', 'souza': 'n/e'},
 'palavra': 'Abusante',
 'reg_amadeu': 'n/e',
 'reg_aurelio': 'n/e',
 'reg_bluteau': 'n/e',
 'reg_cunha': 'n/e',
 'reg_laud': 'n/e',
 'reg_morais': 'n/e',
 'tipo': 'Nm[ADJsing]'}
```

Figura 2. Consultar informações com get (self, chave).

```
tabela.remove('ADOBE')

print(tabela)

    Tamanho total da tabela: 449
    Posição 16: {'palavra': 'Abuscar', 'tipo': 'V', 'ex_uso': 'Uai ô durmi na casa de (R
    Posição 76: {'palavra': 'Abusante', 'tipo': 'Nm[ADJsing]', 'ex_uso': 'Uai ô durmi na
    Posição 181: {'palavra': 'Abisar', 'tipo': 'V', 'ex_uso': 'Tabaiava lá naquela época
    Posição 443: {'palavra': 'Acuier', 'tipo': 'V', 'ex_uso': 'Passado uma semana nós vor
```

Figure 3. Remover informações com remove (self, chave).

```
Tabela Hash!
Léxico do Café

Escolha uma opção:

1. Inserir item
2. Recuperar item
3. Deletar item
4. Mostrar Tabela
5. Teste de performance
5
Mudando tamanho da tabela
Tempo de criacao: 0.5990264415740967
Tamanho da tabela: 1333357
Inserindo 1000000 itens com chaves de 20 caracteres
Tempo de insercao: 4.184812545776367
Média do tempo de insercao: 4.184812545776367e-06
Recuperando 1000000 itens
Tempo de recuperação: 3.4817628860473633
Média do tempo de recuperação: 3.4817628860473632e-06
Deletando 1000000 itens
Tempo de deleção: 3.5478594303131104
Média do tempo de deleção: 3.5478594303131103e-06
Aperte Enter para continuar
```

Figura 4. Medições de velocidade das operações.

Além da interface, também foi criado um arquivo csv contendo o léxico completo e formatado. A Figura 5 mostra uma visualização do léxico utilizando a biblioteca *Pandas*.

	palavra	dicionarizado	categoria_gramatical	idioma_origem	definicao	frase_de_abonacao
0	ABISAR	(A)	[V]	Lat>Port	Dar aviso a, informar, previnir. Variante de a...	ele foi abisado que tava com...mas é num ligav...
1	ABUSANTE	(n/d)	[Adj]	Lat>Port	Que excede o permitido.	Cê faz ôta cumpá (C...) faz ôta dessa fica abu...
2	ABUSCAR	(A)	[V]	obs.	Tratar de trazer ou levar. Variante de buscar.	...fui buscá dispesa é maco maco...ganheí fui ...
3	ACUIER	(n/d)	[V]	(n/e)	Tirar,desprender separando do ramo ou da haste...	...o arroz que tinha ... que tava sem cortá le...
4	ADOBE	(A)	Nm[Ssing]	Ár.	Tijolo feito de argila, seco ou cozido ao sol,...	Fui embora pra casa de mamãe lá ê fez dois côm...
5	ALEMBRAR	(A)	[V]	Lat>Port	Trazer algo à memória, recordar, relembrar. Va...	Ele era fei menina em vida ô num alembro dele ...
6	ALEVANTAR	(A)	[V]	Lat>Port	Colocar ou colocar-se de pé, elevar-se. Varian...	Com deus me deito com deus me alevanto...com a...
7	ALEVAR	(A)	[V]	Lat>Port	Fazer passar de um lugar para outro, carregar,...	É pegô a cabeça dele assim e impurrô ê pra bax...
8	ANDAR IGUAL GALINHA TONTA	(n/d)	[Fras]	(n/e)	Vaguear transtornado, caminhar fora de si.	...nôto dia eu andava o terrero todo iguali ga...
9	ANDAR NA TÁBA DA BERADA	(n/d)	[Fras]	(n/e)	Ver-se em situação de apuro.	Dessa vez o (T...) ...andô na taba da berada d...

Figura 5. Representação dos elementos da tabela pela biblioteca *Pandas.*

Uma vez obtida essa tabela, o próximo passo é disponibilizar esse aplicativo na Web [5], tal que esse possa ser usado por aqueles que, direta ou indiretamente, lidam com os moradores da região da Serra do Cipó.

DISCUSSÃO GERAL

Iniciamos esta apresentação falando da existência de um movimento global de valorização da diversidade cultural. Mais especificamente, esse movimento até onde é possível perceber clama pela inclusão dos grupos socialmente desprestigiados. A valorização da linguagem usada por esses grupos é, por razões óbvias, um dos meios mais eficazes para reconhecer a identidade deles.

De acordo com isso, tecnologias em ILNBD, se pretendem atuar em sintonia com a realidade, devem necessariamente considerar essa exigência. Assim, mais do que evidenciar como a computação pode ser usada em interfaces humano-computador, esta apresentação procurou mostrar que a digitalização do léxico é uma maneira da ILNBD promover a inclusão cultural.

O léxico escolhido foi o glossário do café na Serra do Cipó, o qual apresenta uma visão de mundo típica da região, por meio de lexias ricas em detalhes socioculturais e históricos. Isso pode ser constatado não apenas pelas informações referenciais, mas também e, sobretudo, mediante a materialidade da variante linguística considerada.

A título de exemplo, vejamos o que podemos encontrar analisando um exemplo, tirado de Freitas [18].

BESTIR~BISTIR • (A) • [V] • Lat>Port • O mesmo que vestir. • "Misturô todo mundo as pretinha com nóis né....aí depois que nós bistiu que nós desceu do artar...dom (C...) tava né ê oiô e falô assim "anjo preto é domônio...não beste mais menino preto". (Ent.03, linhas 30 e 31) • (bestir~vestir: caso de degeneração) (p. 226).

Em primeiro lugar, encontramos indicações se as expressões são registradas em dicionário. Nesse caso, as lexias estão dicionarizadas no Aurélio representadas pelo

(A). Isso significa dizer que se trata de expressões há muito tempo presentes na fala dos brasileiros e que, por razões sócio-históricas, se radicaram na região da Serra do Cipó.

Períodos marcantes da história do Brasil tiveram como cenário a Serra do Cipó. A região serviu como via de acesso aos Bandeirantes que partiam de São Paulo em busca de ouro e pedras preciosas. Grandes belezas naturais estão no Parque Nacional da Serra do Cipó, o qual possui um relevo acidentado e altitudes que variam entre 700 e 1700 metros [18].

Depois, encontramos necessariamente referência à origem da palavra, p.ex., "Lat>Port". Esse dado nos informa que o verbete evoluiu do latim e *ipso facto* nos sugere que ele reflete um processo regular de deriva. Assim sendo, a alternância "bestir"-"bistir" não pode ser avaliada como um erro. No final do verbete, encontramos ainda a natureza do metaplasmo, a "denegeração".

Metaplasmo é o termo pelo qual os especialistas denominam a variação observada na forma das palavras ao longo do tempo. Longe de serem considerados erros de grafia ou pronúncia, os metaplasmos são vistos como processos regulares de mudanças, que são determinados por derivas inerentes à própria língua. Um exemplo disso é a alternância sistemática entre "v" e "b" verificada em algumas palavras do português ("vassoura" ~ "bassoura"). Há metaplasmo por permuta, transposição, subtração e aumento que podem ser vistos em Gama e Santos [23].

Finalmente, a abonação nos oferece um exemplo de contexto linguístico, onde o verbete foi registrado durante as entrevistas. Essa informação é ilustrativa da maneira pela qual os moradores locais usam uma variedade do português como meio de comunicação.

Nas últimas décadas temos assistindo a um esforço da Linguística em realizar um estudo objetivo dos fatos verbais. Esse esforço incluiu também a necessidade de se evidenciar a legitimidade das variantes linguísticas. Como dissemos anteriormente, os trabalhos de antropólogos, linguistas e sociolinguísticos foram decisivos para isso. Graças a eles, formas verbais até então marginalizadas, como as línguas ameríndias, os dialetos e as gírias, são consideradas hoje meios sistemáticos e legítimos de comunicação.

Atualmente, com o crescimento das tecnologias de informação, é fundamental que essa diversidade seja também tomada em conta pelas ILNBDs. A divulgação de um glossário como o da cultura rural de uma região é um passo importante nessa direção.

Portanto, isso não pode ser tratado como uma mera curiosidade em relação aos hábitos linguísticos do outro. É preciso o reconhecimento de que as variantes linguísticas estudadas são elas mesmas a cultura das pessoas que as falam. Só assim, com essa mudança de perspectiva diante da linguagem, as interfaces podem criar aplicativos mais sintonizados com a diversidade cultural; e fazer com que os usuários se sintam representados nos meios digitais [5].

CONSIDERAÇÕES FINAIS

Este artigo objetivou mostrar que uma iniciativa de visibilidade sociolinguística é uma atitude de promoção da diversidade cultural entre as tecnologias de informação. Isso vai além do que tratar da indexação em si de um léxico regional.

Com a criação de uma interface para acessar uma tabela hash para a fala rural da Serra do Cipó, enfatizamos que as ILNDBs podem ser perfeitamente instrumentos de inclusão cultural desde que elas possam ter um *parser* que seja capaz de analisar automática e estruturalmente de maneira correta sentenças do português, sem restrições, de uma gama de gêneros textuais, tão vasta quanto possível.

Assim sendo, a proposta apresentada deve ser vista como o início de um trabalho em curso [5]. Um passo além na direção de melhorar a interface requer, ademais, a transformação desse código em um aplicativo manuseável. Algo dessa natureza pode ser obtido por meio de uma interface gráfica amigável, como a oferecida por Moreno [4] que associou à sua hash das ervas o software *Visual Tahs*.

A criação de um aplicativo em nuvem fica como sugestão de melhorias futuras em tabelas hash de variedades linguísticas regionais. Com essa melhoria, e outras mais, acreditamos que o processo de diversidade cultural pode ser satisfatoriamente promovido no domínio das ILNBDs.

Além disso, uma vez disseminado este trabalho, pretende-se disponibilizar o Léxico Rural da Serra do Cipó.

Este trabalho pode ainda incitar outros estudos sobre dialectologia ou dialectometria regional para analisar grandes corpus de mídia social, dado o grande interesse atual nesse tipo de comunicação informal.

REFERÊNCIAS

1. Carolina Manzano. 2020. Diversidade cultural para um desenvolvimento sustentável: contribuições da convenção para a proteção e promoção da diversidade das expressões culturais. In *Anais do XXV Encontro Estadual de História: história, desigualdades e diferenças* (ANPUH'20), Vol. 25, São Paulo.
2. Maria de Lourdes Ferrioti e Dulce M. P. de Camargo. 2008. Diversidade, Educação, cultura e sustentabilidade: relacionando conceitos. *O mundo da Saúde*, 32, 3 (Jul./Set., 2008), 359-366.
3. Alfred V. Aho, Ravi Sethi and Jeffrey D. Ullmann. 1995. *Compiladores:* Princípios, técnicas e ferramentas. Trad. Daniel A. Pinto. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora SA.
4. Fábio C. Moreno. 2017. *Visual Tahs: Ferramenta para analisar a eficácia de buscas das funções hash em um léxico para língua natural.* Dissertação de

Mestrado. Departamento de Computação da Universidade Estadual de Londrina, Londrina.

5. Vitor C. Silva, Cinthyan R. S. C. de Barbosa e Wagner F. Lima. Inclusão Cultural em Interfaces para Banco de Dados: Léxico Rural da Serra do Cipó. In *Anais do XXII Seminário de Estudos sobre Linguagem e Significação* (SELISIGNO'22). UEL, Londrina. 623-637.
6. Cristina M. Fargetti, Clotilde A. A. Murakawa e Odair L. Nadim (Ed.). 2019. *Léxico em foco: dicionários com que sonhamos.* São Paulo: Cultura Acadêmica. Série Trilhas Linguísticas.
7. Bento Dias-da-Silva e Ariani Di Felippo. 2000. *Concepções de Léxico e o Processamento Automático das Línguas Naturais*. Recuperado Outubro 29, 2022 de http://www.gel.hospedagemdesites.ws/estudoslinguisticos/volumes/32/htm/comunica/ci035.htm.
8. William Labov. 1972. *Sociolinguistic pattern*. Philadelphia: University of Pennsylvania Press.
9. Alana R. B. S. Lisboa and Cinthyan R. S. C. de Barbosa. 2013. Lexicon of Orchids. *Procedia Social and Behavioral Sciences*. 95 (Oct., 2013). Elsevier. Alicante, Espanha. 81-88. https://doi.org/10.1016/j.sbspro.2013.10.625
10. Fábio C. Moreno, Cinthyan R. S. C. de Barbosa e Edio R. Manfio. 2021. Tabelas Hash para um Léxico Digital. *Revista de Informática Teórica e Aplicada* (RITA), 28, 2 (Ago., 2021), 26-38. https://doi.org/10.22456/2175-2745.107128
11. Carolinne R. e Faria. 2021. *Ferramenta Carolina para Identificação de Pragas e Doenças na Cultura da Soja utilizando Processamento de Linguagem Natural*. Dissertação de Mestrado. Departamento de Computação da Universidade Estadual de Londrina, Londrina.
12. Aldo M. Paim. 2016. *Inferência de personalidade a partir de textos em português brasileiro utilizando léxicos*. Curitiba: Departamento de Informática da Pontifícia Universidade Católica do Paraná. Dissertação de Mestrado.
13. Erick N. P. de Souza. 2014. *Classificação de relações semânticas abertas baseada em similaridade de estruturas gramaticais na Língua Portuguesa*. Dissertação de Mestrado. Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, Salvador.
14. Juliana G. Greghi. 2002. *Projeto e desenvolvimento de uma base de dados lexicais do português*. Dissertação de Mestrado. Instituto de Ciências Matemática e de Computação da Universidade de São Paulo, São Carlos.
15. Luciene G. F. Marins. 2014. O léxico rural no Brasil Central: designações para "bruaca". *Estudos Linguísticos*, 43, 1 (Jan.-Abr., 2014) , 545-560.
16. Greize A. da Silva e Patrícia A. Borges. 2019. Presença vs ausência de traços de ruralidade no léxico tocantinense. *Revista do Instituto de Estudos Brasileiros*. n.72 (Abril, 2019). 83-105. http://dx.doi.org/10.11606/issn.2316-901X.v0i72p83-105
17. Marco Gonzalez e Vera L. S. de Lima. 2003. Recuperação de informação e Processamento da Linguagem Natural. In *Anais do XXIII Congresso da Sociedade Brasileira de Computação* (SBC'03). SBC, São Paulo, 347–395.
18. Cassiane J. de Freitas. 2012. *Café com quebra torto*: um estudo léxico-cultural da Serra do Cipó/MG. Dissertação de Mestrado. Faculdade de Letras da Universidade Federal de Minas Gerais. Belo Horizonte.
19. O. Yu Mikhailyuk. and H. Ya Pohlod. 2015. The languages we speak affects our perceptions of the world. *Journal of Vasyl Stefanik Precarpathian National University*, 2, 2-3, 36-41. https://doi.org/10.15330/jpnu.2.2.36-41
20. Fábio C. Moreno, Cinthyan R. S. C. de Barbosa e Edio R. Manfio. 2019. Visual Tahs: software para auxiliar o ensino de tabela Hash na disciplina de Estrutura de Dados. In *Anais do XLVI Seminário Integrado de Software e Hardware* (SEMISH'19). SBC, Belém. 33-44. https://doi.org/10.5753/semish.2019
21. Bob Jenkins. 1997. Algorithm alley-what makes one hash function better than another? Bob knows the answer, and he has used his knowledge to design a new hash function that may be better than what you're using now. *Dr Dobb's Journal-Software Tools for the Professional Programmer*, Redwood City, CA, 22, 9 (Sep. 1997), 107–110.
22. Edio R. Manfio, Fábio C. Moreno e Cinthyan R. S. C. de Barbosa. 2014. Professor Tical e AliB: Interação Humano Computador em Diferente Campo. In *Anais do XIX Conferência Internacional sobre Informática na Educação* (TISE'14). SBC, Fortaleza. 782-787.
23. Gislene A. Gama e Leonardo G. dos Santos. 2017. *O internetês como variação na Língua Portuguesa do Brasil.* Recuperado Setembro, 29, 2022 de https://semanaacademica.org.br/system/files/artigos/artigotccgislenedeabreu gama.pdf