Léxico Rural da Serra do Cipó:

um caso de indexação digital de variantes regionais do português

Vitor de Castro Silva
Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Londrina
Londrina, Paraná, Brasil
vitor.castro.silva@uel.br

Cinthyan R. S. C. de Barbosa
Programa de Pós-Graduação em Ciência da Computação
Universidade Estadual de Londrina
Londrina, Paraná, Brasil
cinthyan@uel.br

Wagner Ferreira Lima
Departamento de Letras Vernáculas e Clássicas
Universidade Estadual de Londrina
Londrina, Paraná, Brasil
wflima@uel.br

RESUMO

Como expressão da 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. 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.

ABSTRACT

As the expression of the identity of a people, the lexicon must be studied in its own right. Currently, with the growing wave of valuing differences, the study of lexicons has become even more relevant. The interfaces must take this into account if they want to meet social demands. Therefore, a survey was conducted to promote and recognize rural linguistic variety in interface research. The proposal consisted of indexing the rural vocabulary of Serra do Cipó, metropolitan region of Belo Horizonte-MG, with a Aho hash function obtaining optimization for all operations. Finally, the conclusion is that such a dictionary represents an important means of digital inclusion, and should serve as a model for future work.

KEYWORDS

Digital inclusion, Serra do Cipó, Natural Language Processing.

1 INTRODUÇÃO

A diversidade cultural se dá por meio dos diversos modos das expressões culturais, quaisquer que sejam os meios de tecnologias empregadas. O desenvolvimento acelerado dessas, 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 [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.

A interculturalidade dos meios de comunicação é uma condição histórica irreversível. Não seria possível retroceder a esses movimentos de tempo e espaço que aceleram as trocas culturais e almejar um planeta com sociedades isoladas, fechadas em si mesmas [2].

Porém, ainda hoje há o preconceito linguístico. Embora a língua falada pela grande maioria da população no Brasil seja o português, esse apresenta alto grau de diversidade e de variabilidade, não só por causa da grande extensão territorial do país, mas principalmente por causa da trágica injustiça social [3] que faz do Brasil um país com pior desigualdade de renda e supera só países africanos.

Esses aspectos são ainda mais evidentes, já que é por meio das palavras que grande parte de valores, conhecimento, ideias, tradição e experiências de uma determinada comunidade são expressos. O indivíduo, por meio de sua fala, interage com as outras pessoas de seu grupo e, assim, laços sociais do grupo são fortalecidos [4].

Assim, o objetivo deste trabalho foi mostrar a implementação de um léxico rural da Serra do Cipó/MG, como modelo para projetos que envolvam o Processamento de Linguagem Natural (PLN). Tais termos estão em Freitas [4] e será usada a função hash Aho [5] como método de indexação, devido à sua eficiência [6] e busca dos verbetes do referido léxico.

Este trabalho se justifica por duas exigências sociais atuais [**7**]: 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 inclusão digital, a presente apresentação pretende ser uma contribuição com as pesquisas em Interfaces em Linguagem Natural (ILNs). 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.

As seções 2 e 3 apresentarão, respectivamente, trabalhos correlatos e o léxico-cultural da Serra do Cipó. Na Seção 4 será descrito o método de *hashing* empregado na indexação desse vocabulário. Na Seção 5 os materiais e métodos serão apontados. Na Seção 6 os resultados da implementação serão apresentados. Na Seção 7 uma breve discussão sobre esses resultados será exibida. Por fim, na Seção 8, as considerações finais e trabalhos futuros serão elencados.

2 TRABALHOS CORRELATOS

Para um sistema de PLN, os léxicos estruturados e o formalismo gramatical são centrais. Porém, muitos trabalhos sobre léxicos não dão informações semânticas precisas.

Fargetti, Murakawa e Nadim [8] 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.

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 [9].

Muitos trabalhos sobre léxico têm sido publicados, porém são de conceitos mais específicos, como Léxico das Orquídeas [10], Léxico das Ervas [11] e Léxico das Pragas das Sojas [12].

Quantos aos formalismos gramaticais para ILNs muitos têm sido propostos [13] inclusive para ampla cobertura, capazes de lidar com fenômenos linguísticos complexos e de propor a melhor interação entre sintaxe e semântica.

A Sociolinguística é um ramo da Linguística que se caracteriza por conceber uma língua como uma realidade essencialmente sociocultural [14]. Suas pesquisas são feitas por entrevistas e/ou amostragens. Seu objeto é a língua falada, analisada, descrita em seu contexto social.

A relação entre língua e sociedade permeia os estudos da Sociolinguística que tenta apresentar como variações linguísticas relacionam-se em determinada comunidade e as diferenças existentes na estrutura social dessa comunidade [4].

Ferramentas e métodos de consultas para léxicos digitais são bem escassos para a Língua Portuguesa [15] [16]. 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 [17]. 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 baixos desempenhos nas buscas.

Trabalhos sobre léxicos rurais têm sido estudados. Marins [18] 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 [19] pelos seus traços de ruralidade. Ambos os trabalhos foram apoiados nas seguintes áreas: lexicologia, dialetologia e geolingüística. Essas áreas da linguística fornecem subsídios para o referido trabalho.

Neste artigo o léxico-cultural da Serra do Cipó da região Sudeste será tratado.

3 LÉXICO-CULTURAL DA SERRA DO CIPÓ

O termo léxico, em alguns casos, é utilizado para identificar o componente de um sistema de PLN com informações semânticas e gramaticais sobre itens lexicais. Também é usada a expressão "base de dados lexicais" como sendo uma coleção de informações lexicais, apresentadas em formato estruturado e acessível a sistemas de PLN [20].

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 ILNs **[7]**.

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 [4]. Este artigo apresenta parte de um vocabulário rural da Serra do Cipó, localizada 90 quilômetros de Belo Horizonte na região sul da Cordilheira do Espinhaço no divisor de águas das bacias hidrográficas dos rios São Francisco e Doce.

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 [4] 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.

Variações que as línguas apresentam dependem de fatores socioculturais, tais como classe social, faixa etária, diferenças existentes entre uma região e outra, etc. Essa hipótese se aplica também ao léxico de uma língua. A variação linguística é o resultado da interação dos aspectos sociais e dos aspectos linguísticos.

O grande responsável para ampliação dos estudos relativos à heterogeneidade da língua, onde essa é relacionada a fatos sociais, foi Labov [14], o qual propôs um modelo teórico metodológico que era capaz de sistematizar a "variação natural" da língua falada levando em consideração a relação entre língua e sociedade.

Dentro da linguística, o léxico é visto como o componente no qual mais se notam as influências socioculturais sobre a língua. Esse é variável e seu constante movimento de recriação que amplia o vocabulário de uma língua, o qual é usado por uma comunidade que 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.

Um estudo lexical pode ser lexicológico ou lexicográfico. A diferença é 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 [21].

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 [4] 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.

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 Humano-Computador (IHC) que envolve 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 uma ILN, mas é preciso também conhecer algumas questões técnicas que um projeto desse tipo coloca, a saber: qualquer aplicativo em linguagem natural precisa ser eficiente em seu tempo computacional.

4 FUNÇÃO HASH

Função hash ou *função de espalhamento* tem se mostrado como um dos meios computacionais eficientes na indexação de dados. A seguir será descrita a função hash do Aho, a qual foi escolhida baseando-se nas implementações de 14 funções implementadas por Moreno [6] e discutidas por Moreno, Barbosa e Manfio [22]. Essas funções são baseadas não só em pesquisas em Estrutura de Dados, mas também específicas para Linguagem Natural, projetadas e sugeridas por Jenkins [23] que é um pesquisador da computação e autor de várias funções hash.

O acesso a um léxico de ILN requer um algoritmo eficiente e isso tem a ver basicamente com o tempo e precisão de processamento: acessar um léxico grande de modo rápido e otimizado [6]. Já há algum tempo a indexação por meio da função hash vem satisfazendo esse requerimento. A estrutura de dados decorrente do espalhamento é conhecida por *tabela hash* ou tabela de dispersão.

Um Léxico para Língua Natural é necessário para catalogar as palavras em substantivos, adjetivos, verbos, entre outros, e inserir em uma estrutura de dados que nos retorne o mais rápido possível essa palavra quando necessária e em que contexto está sendo empregada. Para isso contamos com uma função hash que pode ser um diferencial no PLN [6].

Cormen *et al.* [24] definem alguns critérios para uma boa função hash: 1. endereço hash facilmente calculado; 2. fator de carga da tabela hash é elevado para um dado conjunto de chaves; 3. os endereços de hash de um determinado conjunto de chaves são distribuídos uniformemente na tabela hash e uma função hash perfeita se diz ótima quando existe distribuição uniforme de endereços na tabela hash.

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 [6] [11] [23] [25] nessas últimas décadas tem trabalhado nessa direção, dando lugar a diferentes métodos de tratamento de colisões.

A seguir é descrita a função hash Aho que mostrou ser eficiente [11] tendo em conta sua otimização comparada com a das outras funções consideradas [6]. Por isso, optamos por ela nesta interface do léxico rural da Serra do Cipó.

A função Aho é a seguinte [5]:

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 = α*h + (Chave[i])*". Moreno [6] 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.

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.

5 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. Na Tabela 1 a ordenação das propriedades dos verbetes, tal como encontrada em Freitas [4], seguida da descrição da construção da estrutura dos dados.

LEXIA; dicionarização; categoria gramatical; origem; definição; abonação; número da entrevista e linha do corpus; (eventualmente) variação linguística. Por exemplo: "ALEVANTAR • (A) • [V] • Lat>Port • Colocar ou colocar-se de pé, elevar-se. Variante de levantar. • 'Com deus me deito com deus me alevanto...com a graça divina e o sinhô isprito santo' (Ent.01, linha 333) • (alevantar~levantar: caso de prótese)" [4].

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 [6], 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 função hash pelo primo mais próximo do tamanho da tabela (*m*). Ou seja, h mod m. Finalmente, o resto desse valor pelo tamanho real da tabela é o hash da chave de entrada.

Tabela 1: Abreviações e convenções do glossário

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	

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.

6 RESULTADOS

Com uma quantidade dos verbetes indexados (430 expressões), foi possível ter uma noção da eficiência da implementação da referida tabela hash. Além disso, com base em pesquisas prévias (como o aplicativo Professor Tical de Manfio, Moreno e Barbosa

[25]), 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, 3, respectivamente) permitindo assim que uma interface de baixo custo computacional com o usuário seja possível. O acesso a essas operações se deu por meio de um menu de opções de comandos (Figura 1), o qual inclui a opção para comprovar a velocidade de tais operações (Figura 4).

```
Escolha uma opção:

1. Inserir item
2. Recuperar item
3. Deletar item
4: Mostrar Tabela
5: Teste de performance
6: Sair
1
Digite a chave do item que será inserido
teste
Digite a chave e o valor da chave no dicionário, ou digite 'q!' finalizar
teste
teste
Digite a chave e o valor da chave no dicionário, ou digite 'q!' finalizar
q!
Item {'teste': 'teste'} inserido na tabela
Aperte Enter para continuar
```

Figura 1: Menu com opções de manipulação dos dados e o exemplo fictício de inserção do item "teste"

A busca e a remoção de itens ocorrem por meio das opões 2 e 3, respectivamente (Figuras 2 e 3):

```
Digite a chave do item que será recuperado
alembrar
Item recuperado! {'palavra': 'ALEMBRAR', 'dicionarizado': '(A)',
 memória, recordar, relembrar. Variante de lembrar', 'frase_de_ab
 muito fiuzim... (Ent. 06, linha 375)'}
```

Figura 2: Busca de itens (opção 2)

```
3
Digite a chave do item que será deletado
carreta
Item removido!
Aperte Enter para continuar
```

Figura 3: Remoção de itens (opção 3)

As opões 4 e 5 servem, nessa ordem, para consultar os itens da tabela hash e para comprovar a velocidade das operações realizadas. A Figura 4 estima em segundos o tempo dispendido na construção e manipulação de uma tabela contendo 1.333.357 posições.

```
5
Mudando tamanho da tabela
Tempo de criacao: 0.8435642719268799
Tamanho da tabela: 1333357
Inserindo 1000000 itens com chaves de 20 caracteres
Tempo de insercao: 6.279748916625977
Média do tempo de insercao: 6.279748916625976e-06
Recuperando 1000000 itens
Tempo de recuperação: 6.139173984527588
Média do tempo de recuperação: 6.139173984527588e-06
Deletando 1000000 itens
Tempo de deleção: 5.2487897872924805
Média do tempo de deleção: 5.2487897872924804e-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
ABISAR	(A)	[V]	Lat>Port	Dar aviso a, informar, previnir. Variante de avisar.	ele foi abisado que tava com...mas é num ligava pra coisa né pra pressão arta né. (Ent.07,linha 180)
ABUSANTE	(n/d)	[Adj]	Lat>Port	Que excede o permitido.	Cê faz ôta cumpá (C...) faz ôta dessa fica abusano cê é muito abusante. (Ent. 10, linha 227)
ABUSCAR	(A)	[V]	obs.	Tratar de trazer ou levar. Variante de buscar.	...fui buscá dispesa é maco maco...ganhei fui abuscá...busquei...ô truxe um sacão assim ó tudo cheio (Ent.04,linha 152)
ACUIER	(n/d)	[V]	(n/e)	Tirar,despren der separando do ramo ou da haste; apanhar.	...o arroz que tinha ... que tava sem cortá levantô e nózi ... nós consiguiu acuiê. (Ent.07,linha 28)

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

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

7 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 ILN, 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 IHC, este trabalho procurou

mostrar que a digitalização do léxico é uma maneira de 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 a busca apresenta na Figura 2:

{'palavra': 'ALEMBRAR', 'dicionarizado': '(A)', 'categoria_gramatical': '[V]', 'idioma_origem': 'Lat>Port', 'definicao': 'Trazer algo à memória, recordar, relembrar. Variante de lembrar', 'frase_de_abonacao': 'Ele era fei menina em vida ô num alembro dele não mais diz que ê era fei demais muito fiuzim... (Ent. 06, linha 375)'}

Em primeiro lugar, encontramos entre chaves indicações de se as expressões são registradas em dicionário. No exemplo, o item está dicionarizado no Aurélio, representado por (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 [4].

Depois, encontramos referência à classe gramatical do léxico, vale dizer, um verbo [V]. E a seguir temos a indicação de que esse item evoluiu do latim para o português: “Lat>Port”. A busca desse item confere, além disso, informações sobre o significado de “ALEMBRAR”, tais como definição (“Trazer algo à memória, recordar, relembrar.”) e uso (“variante de lembrar”, esse último considerado a forma padrão no português). A abonação também é apresentada, extraída da linha 375, entrevista 6: 'Ele era fei menina em vida ô num alembro dele não mais diz que ê era fei demais muito fiuzim... (Ent. 06, linha 375)'; ilustrando a maneira pela qual os moradores locais usam uma variedade do português como meio de comunicação.

O fato de o lexema “ALEMBRAR” estar dicionarizado indica a presença de processos de variação linguística radicada no português brasileiro. Com efeito, há muito o verbo “alembrar” coexiste com a forma padrão “lembrar”. O acréscimo de um segmento no início de uma palavra é um processo de variação fonológica conhecido como prótese [26]. (Outros exemplos, “soar>assoar”; “voar>avoar”, “renegar>arrenegar” etc.) Esse processo, conquanto desprestigiado socialmente, reflete na realidade um mecanismo regular de variação verbal. Portanto, ele deve ser tratado com respeito e suas manifestações vistas como a expressão de uma comunidade.

Temos assistido a um esforço da Linguística nas últimas décadas 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 ILNs. A divulgação de um glossário como o da cultura rural de uma região é um passo importante nessa direção. Contudo, até onde foi possível evidenciar, a indexação e digitalização de léxicos regionais deve contar com métodos de hash eficientes, como a função Aho descrita. Conquanto o léxico da Serra do Cipó constitua uma amostra relativamente pequena, nossa tabela é, como vimos (Figura 4), capaz de indexar uma quantidade muito maior de itens em pouquíssimo tempo de processamento.

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 **[7]**.

8 CONSIDERAÇÕES FINAIS

Conferir visibilidade digital à variedade linguística de uma comunidade rural é, como defendemos, uma forma de promover a diversidade cultural entre as tecnologias de informação. Nesse sentido, as ILNs podem ser perfeitamente instrumentos de inclusão cultural, contanto que elas possam ter um *parser* que seja capaz de analisar automática e estruturalmente de maneira correta sentenças do português que possam ocorrer naturalmente, sem restrições, de uma gama de gêneros textuais, tão vasta quanto possível. A interface para acessar uma tabela hash para a fala rural da Serra do Cipó aponta nessa direção. Ela é capaz de indexar uma grande quantidade de itens e propiciar buscas eficientes de informações sociolinguísticas.

Vale lembrar que a tabela hash apresentada é o início de um trabalho em curso **[7]**. Um passo além pode ser dado na direção de melhorar a interface, como transformar esse código em um aplicativo manuseável. Uma interface gráfica amigável, como aquela oferecida por Moreno [6] que associou à sua hash das ervas o software *Visual Tahs*, pode ser uma possível solução para essa melhoria. Com ela, acreditamos que a diversidade cultural pode ser satisfatoriamente inserida no domínio das ILNs.

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

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] Marcos Bagno. 2001. *Preconceito Lingüístico*: o que é, como se faz. São Paulo: Loyola.

[4] Cassiane J. de Freitas. 2012. *Café com quebra torto*: um estudo léxico-cultural da Serra do Cipó/MG. Belo Horizonte: Faculdade de Letras da Universidade Federal de Minas Gerais. Dissertação de Mestrado.

[5] 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.

[6] Fábio C. Moreno. 2017. *Visual Tahs: Ferramenta para analisar a eficácia de buscas das funções hash em um léxico para língua natural*. Londrina: Departamento de Computação da Universidade Estadual de Londrina. Dissertação de Mestrado.

[7] Vitor C. Silva, Cinthyan R. S. C. de Barbosa e Wagner F. Lima. 2022. Inclusão Cultural em Interfaces para Banco de Dados: Léxico Rural da Serra do Cipó. In *Anais do XII Seminário de Estudos sobre Linguagem e Significação (SELISIGNO'22) – Processamento de Linguagem Natural (PLN) e suas Implicações para o Entendimento da Língua Materna.* UEL, Londrina.

[8] 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.

[9] Bento Dias-da-Silva e Ariani Di Felippo. 2000. *Concepções de Léxico e o Processamento Automático das Línguas Naturais*. http://www.gel.hospedagemdesites.ws/estudoslinguisticos/volumes/32/htm/comunica/ci035.htm

[10] 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. DOI: https://doi.org/10.1016/j.sbspro.2013.10.625

[11] 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. DOI: https://doi.org/10.22456/2175-2745.107128

[12] Carolinne R. e Faria. 2021. *Ferramenta Carolina para Identificação de Pragas e Doenças na Cultura da Soja utilizando Processamento de Linguagem Natural*. Londrina: Departamento de Computação da Universidade Estadual de Londrina. Dissertação de Mestrado.

[13] Cinthyan R. S. C. de Barbosa. 1998. *Gramáticas para Consultas Radiológicas em Língua Portuguesa*. Porto Alegre: Instituto de Informática da Universidade Federal do Rio Grande do Sul. Dissertação de Mestrado.

[14] William Labov. 1972. *Sociolinguistic pattern*. Philadelphia: University of Pennsylvania Press.

[15] 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.

[16] Erick N. P. de Souza. 2014. *Classificação de relações semânticas abertas baseada em similaridade de estruturas gramaticais na Língua Portuguesa*. Salvador: Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia. Dissertação de Mestrado.

[17] Juliana G. Greghi. 2002. *Projeto e desenvolvimento de uma base de dados lexicais do português*. São Carlos: Instituto de Ciências Matemática e de Computação da Universidade de São Paulo. Dissertação de Mestrado.

[18] Luciene G. F. Marins. 2014. O léxico rural no Brasil Central: designações para "bruaca". *Estudos Linguísticos*, 43, 1 (Jan-Abr., 2014) , 545-560.

[19] 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. DOI: http://dx.doi.org/10.11606/issn.2316-901X.v0i72p83-105

[20] 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, Campinas, 347–395.

[21] 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. DOI: https://doi.org/10.15330/jpnu.2.2.36-41

[22] 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. DOI: https://doi.org/10.5753/semish.2019

[23] 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.

[24] Thomas H. Cormen. Charles E. Leiserson, Ronald L. Rivest and Clifford Stein. 2012. Algoritmos: teoria e prática. Rio de Janeiro: LTC.

[25] 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.

[26] Gislene A. Gama e Leonardo G. dos Santos. 2017. O internetês como variação na Língua Portuguesa do Brasil. https://semanaacademica.org.br/system/files/artigos/artigotccgislenedeabreu gama.pdf