laborhistórico

Revista LaborHistórico
ISSN 2359-6910
v.9, n.1, e52576, 2023
DOI: 10.24206/lh.v9i1e52576

Artigo Original

MICONTES – minicorpus de contos de encantamento do Serro em Python

MICONTES - minicorpus of enchantment tales from Serro in Python

MICONTES - minicorpus de cuentos de encanto de Serro en Python

José Carlos Costa
Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil.
E-mail: carlosjuniorcosta1@gmail.com

Valdinei Pedro Sales Vieira
Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil.
E-mail: pedro.0688@yahoo.com.br

Recebido: 20/05/2022
Aceito: 01/06/2023

Como citar:
COSTA, José Carlos; VIEIRA, Valdinei Pedro Sales. MICONTES - minicorpus de contos de encantamento do Serro em Python. *Revista LaborHistórico*, v.9, n.1, e52576, 2023. doi: https://doi.org/10.24206/lh.v9i1e52576

Resumo

O objetivo deste trabalho é apresentar o Minicorpus de Contos de Encantamento do Serro (MICONTES) em ambiente Python, assim como as tecnologias já disponibilizadas para explorar essas narrativas orais. Metodologicamente, foram criados dois programas principais que normalizam o português semiortográfico da transcrição, etiquetam por classes de palavras, fornecem seus lemas e detalhes morfológicos, contam fenômenos e disfluências, entre outras implementações. Portanto, fornecemos tecnologia para estudiosos que queiram se debruçar sobre essa importante manifestação cultural, ainda que nada saibam de programação.

Palavras-chave

Contos de Encantamento, Processamento de Linguagem Natural, Literatura oral, Linguística Aplicada. Python.

Abstract

The aim of this paper is to present the Minicorpus de Contos de Encantamento do Serro (MICONTES) in Python environment, as well as applications already available to explore these oral narratives. For this purpose it was built two main programs that normalize Portuguese spelling out of transcriptions conventions, count speech disfluencies, tag part of speech, detail word frequency and morphology, lemmatize lexical words, among other implementations. Therefore, we offer technology to shed light on a lot of linguistic phenomena that are present in theses oral tales, even to the ones that are unfamiliar with coding in Python.

Keywords

Enchanting tales, Natural Language Processing, Oral literature, Applied linguistics, Python.

Resumen

Este artículo presenta el Minicorpus de Contos de Encantamiento do Serro (MICONTES) en Python, así como aplicaciones ya disponibles para explorarlo. Para ello, hemos hecho dos programas que normalizan el portugués semiortográfico de las transcripciones, cuentan problemas del habla, hacen su etiquetado gramatical y lematización, detallan su frecuencia de palabras y morfología, entre otras implementaciones. Así, ofrecemos datos y tecnología para que otros investigadores puedan acceder y estudiar estas narrativas orales, aunque nada sepan de programación en Python.

Palabras clave

Cuentos de encantamiento, Procesamiento del Lenguaje Natural, Literatura oral, Linguística Aplicada. Python.

Este trabalho apresenta o Minicorpus de Contos de Encantamento do Serro - MICONTES (VIEIRA, 2018) em ambiente Python, bem como as transcrições e ferramentas computacionais já disponibilizadas para explorá-lo[1].

Na seção 2, são apresentados um panorama teórico sobre contos de encantamento, assim como a arquitetura geral do minicorpus e os procedimentos utilizados para gravá-lo. Na seção 3, encontra-se a metodologia empregada para transpor o

[1] O Micontes já está disponível no link a seguir, no qual também são atualizados os códigos e apresentadas novas ferramentas periodicamente: <https://github.com/carlosjuniorcosta1/Micontes_minicorpus_contos_encantamento>.

minicorpus de Vieira (2018) para ambiente Python, tais como normalização ortográfica e etiquetagem de classes gramaticais, assim como são apresentados detalhes de sua arquitetura e como funcionam os programas criados para contemplá-lo.

Contos de encantamento

O termo "narrativa oral" designa um conjunto amplo de práticas sociais que se materializam textualmente como causos, contos, anedotas, adivinhas, mitos, lendas, entre outros gêneros. Esse tipo de texto é marcado pela dinamicidade da transmissão oral no decorrer de gerações, de modo que não é possível compactá-los pelo postulado de uma origem ou de uma autoria única.

Segundo Cascudo, os contos de encantamento pertencem a uma subclassificação dos contos e são caracterizados por elementos específicos. O encantamento interfere na narrativa em favor do herói que, geralmente, é caracterizado pelo mais moço ou pelo terceiro dos filhos. O herói dos contos de encantamento, como aponta Cascudo (2006, p. 287), é o "indicado para a mais lógica de todas as derrotas", por possuir características como "o doente", "o mais fraco", "o mais novo", disto decorre a necessidade do encantamento para que possa vencer o bruxo, o velho ou a velha, ou o feiticeiro que são personificados como vilões. Em outro momento, Cascudo completa essa definição ao reiterar que os contos de encantamento são caracterizados, principalmente, pela presença do elemento mágico, do sobrenatural, do encantamento, dos dons, dos amuletos, das varinhas de condão e das virtudes acima da medida humana e natural.

Vieira, criador do minicorpus (cf. seção a seguir, para detalhes) acrescenta outros elementos como pertencentes a esse gênero textual. Por meio da análise do Tópico Discursivo e da Perspectiva Textual Interativa, Vieira defende que os contos de encantamento possuíam, de modo intercambiável, pelo menos onze elementos: o Envio; a Tarefa; a Hospitalidade; o Objeto Mágico; o Sobrenatural; o Auxílio Externo; o Elemento Simbólico; o Desconhecimento; o Reconhecimento; o Grotesco; e a Finalização. Durante a contação, esses elementos podiam não ser percebidos conscientemente, mas acreditou-se serem eles os responsáveis pelo encantamento provocado na plateia. De modo mais sintético, o pesquisador concluiu que os contos de encantamento podem ser definidos, relativamente, como uma narrativa oral em que o objeto mágico e os seres sobrenaturais, responsáveis pelo encantamento, servem de auxílio ao herói, que sempre parte, enfrenta adversidades terrenas e extraterrenas, cumpre uma missão, e regressa ao lar.

Dados sociolinguísticos do MICONTES

Acompanhando alguns trabalhos com narrativas orais no contexto brasileiro (FERREIRA, 2010; LIMA, 2003; MARCHUSCHI, 2008; PEREIRA, 1996; WEITZEL, 1995; dentre outros), em específico na abordagem do conto popular, a pesquisa de Vieira (2018) adentra pelos caminhos do Vale do Jequitinhonha, mais precisamente do Serro – MG, para registrar narrativas orais e contos de encantamento do local. Segundo Vieira (2018), existe uma escassez de amostras naquela região e poucos registros impressos e investigações, sobretudo de natureza linguística, a respeito de narrativas orais provenientes do Alto Jequitinhonha.

Em seu trabalho, Vieira (2018/) analisa as narrativas de um único contador: o Sr. José Antônio, à época com 76 anos, provavelmente o último contador vivo de Pedro Lessa, distrito do município de Serro. O minicorpus completo soma sessenta e seis narrativas orais – que ainda estão sendo transcritas – com os mais diversos temas. Trata-se de um conjunto de dados pessoais do pesquisador, que coletou durante muitos anos de sua vida esse quantitativo de narrativas, visto que o contador é um membro da família. As gravações foram realizadas na casa do Sr. José Antônio, localizada na área rural de Serro. Assim, descreve Vieira (2018):

> Seu Antônio se ajeitava próximo a mim e então começava a indagar sobre qual história eu queria ouvir. Contava cerca de dez histórias a cada encontro. Era um tanto considerável, visto que a sua voz não alcançava mais a mesma força de outrora, e a sua memória demorava mais para dar continuidade ou interligar os trechos das narrativas. Tossia e raspava a garganta. Recuperava a voz. Coçava os cabelos brancos. Arranjava as palavras. Quando eu percebia que ele estava se esforçando muito, agradecia e dizia que voltaria a ouvi-lo na semana seguinte. Foi o que fiz ao longo de três meses. Vale destacar que, embora o uso dos aparelhos celulares fosse algo já presenciado por Seu Antônio, o uso do gravador lhe causou estranhamento: "— deixa eu ouvi!". Era o que dizia após cada gravação. E ria quando ouvia a si mesmo, a sua voz, saindo de um objeto tão pequeno... E passávamos para outro conto. (VIEIRA, 2018, p. 37)

A amostra que já foi transcrita por Vieira é formada por doze contos orais de encantamento, visto que as narrativas do MICONTES apresentavam outros gêneros, como anedotas, adivinhas, fábulas e causos. Uma sinopse de cada texto dessa amostra pode ser lida na seção 3.2.

Nas primeiras transcrições, registradas na pesquisa de Vieira (2018), percebeu-se que era necessário manter, dentro das limitações, as características do dialeto do contador. A tentativa do pesquisador foi de preservar não só as propriedades da

fala do Sr. José Antônio em uso real, mas também resguardar o próprio processo de estruturação e produção dos contos populares. Ao conservar determinadas marcas da fala, compreendeu-se que os contos, por meio da voz do contador, estavam estritamente vinculados a determinado grupo social, com particularidades linguísticas, culturais, históricas e sociais que são únicas, que denotam traços identitários, ainda que bastante delimitados, dos moradores do Alto do Jequitinhonha.

Da grande riqueza de detalhes da transcrição emerge um português semiortográfico com diversos símbolos que registram fenômenos variados da fala, como pausas, alongamentos e palavras interrompidas. Se por um lado esse modelo de transcrição permite visualizar onde esses fenômenos ocorrem e contabilizá-los mais facilmente, por outro dificulta enormemente a leitura e processamento computacional.

Essa dificuldade pode ser vista nos etiquetadores de classes de palavras como o Mac-Morpho (ALUÍSIO, 2003), que são treinados em textos com português, em sua maioria, padrão. Além disso, operações como cálculo de similaridade lexical, por exemplo, poderiam ser prejudicadas, visto que palavras diferentes em português semiortográfico representariam apenas uma no português padrão. Por exemplo, *para, pa'* e *pra* seriam codificadas de forma diferente, mas se referem unicamente à preposição *para*. Até mesmo operações mais simples como contagem da frequência de palavras seria prejudicada, visto que cada palavra seria contada de forma diferente – e isso nem sempre é de interesse do linguista ou do programador.

Por isso, foram necessárias diversas codificações para realizar a etiquetagem por classes de palavras e fazer as mais variadas contagens nas transcrições, de forma a fornecer ao usuário acesso rápido e eficiente a muitas variáveis que lançam luz sobre o estudo dos contos. Esses procedimentos, bem como a arquitetura do minicorpus em ambiente Python são discutidos ao longo da seção 3 a seguir.

Metodologia

Estrutura geral das transcrições

O MICONTES possui 12 áudios transcritos que contabilizam 13.329 palavras. As transcrições possuem extensão bastante variada, com contos que têm de 438 a 1827 palavras, conforme é possível ver na Figura 1.

Esses contos têm a duração de 4 a 14 minutos. Tal como sua quantidade de palavras, sua duração também é bastante heterogênea (Figura 2).

Na subseção a seguir, é possível conhecer um pouco de cada um desses contos por meio de uma breve sinopse.

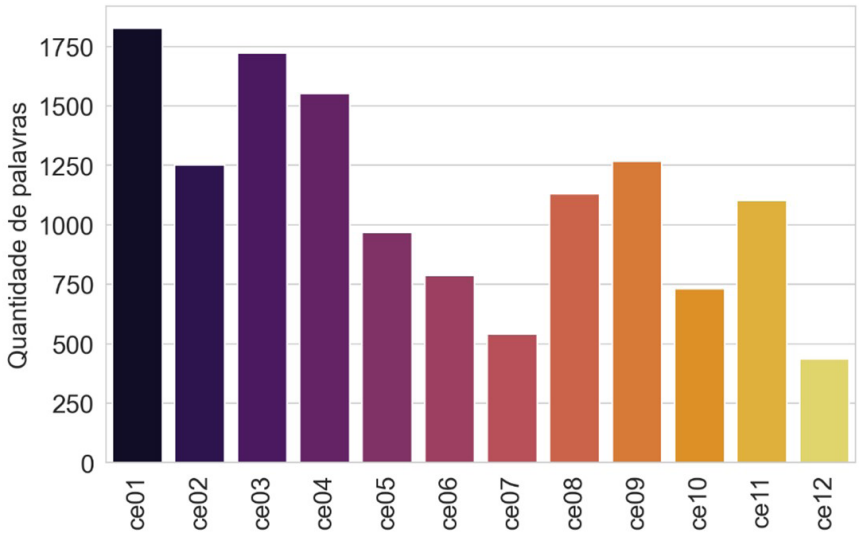

Figura 1. Quantidade de palavras por transcrição em Micontes

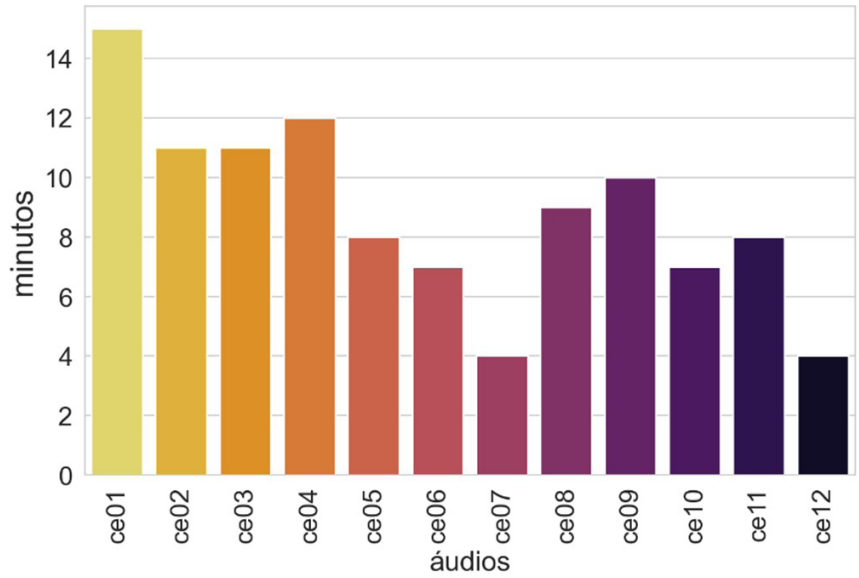

Figura 2. Duração dos áudios do minicorpus

Sinopse dos contos de encantamento

CE-01 Maria Caçula, Maria do Meio e Maria Mais Velha

Três irmãs são raptadas pelo Rei dos Gaviões, Rei dos Bois e Rei dos Peixes. Joãozinho, o filho caçula, desejando conhecê-las, enfrenta seres sobrenaturais e estende sua aventura na busca pelo amor de Maria Bonita, a filha do rei.

CE-02 O gigante

Sendo chantageado pelos irmãos e tendo que cumprir vários desafios impostos pelo Rei, inclusive o de capturar um gigante, Joãozinho ganha a mão da princesa em casamento.

CE-03 Os dois meninos gêmeos

Envolvida em um forte feitiço, uma princesa permanece emudecida. Até que um dos meninos gêmeos vai até a Ilha dos Bichos e corta três pontas da língua da Serpente para livrá-la da maldição. Entretanto, em uma nova inventiva, este menino gêmeo fica preso no Lugar que Vai e Não Volta, sendo ajudado pela premonição de perigo do seu irmão gêmeo que parte para salvá-lo.

CE-04 Adão e Diomara na Terra de Era

Com a ajuda da Mãe do Sol, da Mãe da Lua e da Mãe do Vento, Diomara enfrenta uma longa e perigosa jornada para encontrar Adão, seu prometido amor. Mas ao chegar à Terra de Era, uma triste descoberta: Adão estava prestes a se casar com uma princesa.

CE-05 Joãozinho e Maria

No desespero causado pela pobreza, um pai abandona os dois filhos na mata: Joãozinho e Maria. Ao tentar achar um lugar seguro, os dois são aprisionados por uma velha que os alimenta como crias de engorda, na intenção de devorá-los mais tarde.

CE-06 Joãozinho e Diomara

Contrariando o desejo do Rei e da Rainha, Diomara foge para se casar com Joãozinho. Entretanto, ao saber da fuga, o Rei inicia uma perseguição ao casal, que só consegue se desvencilhar pela ajuda de elementos mágicos.

CE-07 Joãozinho e Estaleiro Estalão

Depois de muitas tentativas, Joãozinho captura uma pomba juriti encantada que se transforma em princesa. Mas Mula Torta, com inveja, espeta um alfinete na pombinha que parte para longe.

CE-08 O chicotinho

Diferente dos irmãos, Joãozinho parte de casa pedindo muita benção e pouco dinheiro. Seus irmãos gastam todo o dinheiro no jogo e vão presos. No meio do caminho uma velha presenteia Joãozinho com objetos mágicos e ele segue para resgatar os irmãos.

CE-09 O pé de Feijão

Jaques troca a vaca da família por três sementes de feijão. Ao serem jogados pela janela os grãos crescem sem parar até atingir a casa do gigante. Então, Jaques sobe pelo pé de feijão e começa a furtar objetos mágicos do gigante. Na última visita, Jaques é descoberto e inicia-se uma perseguição.

CE-10 Lâmpada velha por lâmpada nova

Sem saber que era uma lâmpada mágica, a princesa troca a lâmpada velha de Joãozinho por uma lâmpada nova. Joãozinho recorre a ajuda de animais encantados para recuperar sua lâmpada mágica e também sua amada, que agora está sob posse do ladrão.

CE-11 Joãozinho Borralheiro

Com a ajuda dos cavalos encantados Mimoso, Corta Vento e Acode Com Tempo, Joãozinho, que até então só fica esquentando fogo, consegue ir à cidade para enfrentar o maior desafio de todos: no alto do trono, quem conseguir tirar o anel do dedo da princesa irá casar-se com ela.

CE-12 Moça da Figueira

Uma madrasta malvada manda enterrar a princesa no pé de uma figueira. Tempos depois, quando alguns homens fazem a capina do quintal, uma voz de menina entoa uma cantiga e então eles rapidamente avisam o Rei.

Onde baixar o minicorpus

O minicorpus completo pode ser baixado diretamente pelo GitHub[2], conhecida rede social utilizada por desenvolvedores em todo o mundo. No site, são fornecidos tanto o *corpus* transcrito quanto seus áudios. As transcrições estão em uma tabela em formato .csv, que é bastante utilizado para análise de dados e programação, e xlsx, a qual pode ser aberta diretamente no Excel, sem necessidade de conversão. Nessas tabelas, são disponibilizadas todas as transcrições do minicorpus e diversas informações, conforme discutido ao longo deste trabalho.

A opção pelo GitHub se justifica devido à facilidade de atualização dos repositórios, gratuidade, revisão de código e contato com a comunidade Python em torno do mundo, a qual ajuda a melhorar progressivamente o trabalho. À medida que outros áudios vão sendo transcritos e revisados, novos arquivos com novos contos são publicados no repositório.

Onde executar os programas

Os dois programas oferecidos são executados no Google Colab ou em qualquer ambiente Python, tal como o Spyder. No caso do Google Colab, o usuário deve clicar nos links disponibilizados no arquivo README.txt de cada programa no GitHub e, após ser direcionado ao Notebook correspondente ao programa, fazer upload dos arquivos solicitados e clicar no executar. Esse procedimento dispensa conhecimentos de programação. Caso o usuário tenha algum conhecimento de programação, pode utilizar outro ambiente de execução Python.

[2] Ver nota 1.

Programa 1

O Programa 1 fornece as seguintes informações:

I. Normalização ortográfica do minicorpus
II. Etiquetagem por classes gramaticais
III. Quantidade de palavras por arquivo
IV. Palavras mais frequentes
V. Palavras com redução segmental
VI. Truncamentos – palavras interrompidas
VII. Alongamentos vocálicos e consonantais
VIII. Palavras diante de pausas
IX. Classes de palavras diante de pausas
X. Quantidade de palavras por transcrição

Entre esses itens, destacamos a normalização ortográfica, que transforma português não padrão em padrão, assim como limpa a transcrição de caracteres metalinguísticos, e a etiquetagem por classes de palavras, a partir da qual são realizadas diversas contagens e verificações sobre o léxico. Esses procedimentos são discutidos a seguir.

Normalização ortográfica

Este procedimento transforma o português não padrão das transcrições de Vieira (2018) em português padrão. Procedimento semelhante também foi executado em transcrições que seguiam outros critérios e representavam outros corpora orais do português (COSTA JÚNIOR, 2022). A normalização viabiliza a etiquetagem de classes de palavras e outros recursos, a serem discutidos nas próximas seções, como lematização e detalhamento morfológico dessas classes por meios automáticos. Desse modo, é possível registrar fenômenos importantes da fala nas transcrições, tais como pausas, palavras interrompidas ou alongamentos vocálicos, sem inviabilizar o estudo do léxico que requisite ortografia tradicional.

Trata-se de um procedimento complexo que envolve tanto a limpeza de fenômenos fonéticos registrados na transcrição quanto substituição de palavras e expressões previamente cadastradas pelas formas do português padrão. Além de normalizar as transcrições já existentes, isso permite que novas transcrições sejam normalizadas – se seguidos os mesmos critérios que Vieira (2018) de forma automática.

O Quadro 1 abaixo ilustra esse procedimento.

Possíveis inadequações de concordância nominal e verbal são mantidas. Por exemplo: *tinha três filho* continua *três filho* e não *três filhos*. Já *um dos mais véio falo(u)* se transforma em *um dos mais velho falou*. Isso foi feito por dois motivos. O primeiro, para preservar ao máximo fenômenos da fala e não tentar corrigi-la baseado em convenções normatizadas da escrita ou da fala formal, desde que esses fenômenos

Quadro 1. Processo de normalização ortográfica

Transcrição	Transcrição normalizada
tinha um home(m) que ela era casado tinha três fi(lh)o... tinha Mané, Pedro e João... aí um dos mais véio falo(u) assim – ah:: meu pai eu vo(u) sai(r) pra trabaiá/ caça(r) emprego! aí... – ah:: num vai não meu fi::(lh)o! aí... aí – que que cê que(r) muito dinhe(i)ro e po(u)ca benção ou muita benção e po(u)co dinhe(i)ro?	tinha um homem que ela era casado tinha três filho, tinha Mané, Pedro e João, aí um dos mais velho falou assim – ah meu pai eu vou sair para caçar emprego! aí, – ah num vai não meu filho! aí, aí – que que você quer muito dinheiro e pouca benção ou muita benção e pouco dinheiro?

não impactassem negativamente na etiquetagem de classes de palavras. O segundo motivo de manter ocasionais inadequações da fala, como *filho* no lugar de *filhos* e *velho* no lugar de *velhos*, nos exemplos discutidos, se justifica pelo fato de que nem todo *filho* deveria ser substituído por *filhos*, da mesma forma que para o adjetivo *velho*. Consequentemente, o código fica mais viável de ser realizado e disponibilizado para os usuários.

Depois disso, é possível partir para a etiquetagem de classes gramaticais nos contos de encantamento, explicitada na seção a seguir.

Etiquetagem por classes de palavras

A etiquetagem é feita após o treinamento de um etiquetador do tipo Brill no Mac-Morpho (ALUÍSIO *et al*, 2003), Minicorpus gratuito, disponível na NLTK, com 51.397 sentenças já etiquetadas com 1.170.095 palavras. Parte desse código é inspirado no de Mateus Inoue[3] e na ampla documentação da NLTK[4] desse processo, na qual o código de Inoue também foi baseada. A diferença principal é no tratamento extensivo com o Pandas e expressões regulares após a etiquetagem, de acordo com as necessidades das transcrições de nosso minicorpus. Esse procedimento se deve à necessidade de diminuir os erros de etiquetagem, visto que o Mac-Morpho, usado como corpus de treinamento, é majoritariamente baseado em dados de escrita e não de fala, como nosso corpus requisitaria.

Os etiquetadores utilizados são o DefaultTagger, o AffixTagger; UnigramTagger, BigramTagger, TrigramTagger e BrillTagger. Apesar de poder ser utilizado o RegexTagger, de expressões regulares, como backoff dos outros, a eficácia correção da etiquetagem por meio do módulo re no Pandas se mostrou muito mais eficiente. Por isso, não é recomendado utilizar esse etiquetador de classes de palavras fora do conjunto de códigos do Programa 1, uma vez que é o Pandas e a mineração de texto, presentes no programa e não no etiquetador, que fazem diversas correções posteriores ao processo que envolve a NLTK aqui descrito.

[3] https://github.com/inoueMashuu.

[4] https://www.nltk.org/book/ch05.html.

Cada um desses etiquetadores tem a função de fornecer uma base para o etiquetador anterior e melhorar sua acurácia, a partir de cálculos de probabilidade que envolvem procedimentos diferentes. Primeiramente, é atribuído um etiquetador default, chamado de DefaultTagger. Esse etiquetador recebe a etiqueta provavelmente mais frequente no *corpus* que, no caso do Mac-Morpho, é o substantivo (N). Para nosso minicorpus, também foi utilizado esse procedimento e, quando todos os etiquetadores que o sucedem falham em reconhecer a classe de palavra em questão, a esta é atribuído um "N", de substantivo, pois a probabilidade de ser um substantivo é maior do que todas as outras. A acurácia desse etiquetador é de apenas 19% de acerto.

Posteriormente, foi construído um AffixTagger, que foi programado para verificar a periferia direita das palavras, uma vez que, em português, o sufixo pode ajudar na identificação de uma classe de palavra. De fato, a acurácia do modelo aumenta, mas ainda é bastante baixo, com 36% de acerto.

Após esse etiquetador, foi treinado um UnigramTagger, que basicamente é o sentido dicionarizado da palavra. Há um salto de acurácia bastante importante, para 83,75% de acerto. A partir deste ponto, os ganhos obtidos com o BigramTagger, que combina duas palavras, e com o TrigramTagger, que combina três, são mínimos, com 85,22% e 85,24%, respectivamente. É com o BrillTagger, que calcula regras funcionais e úteis a etiquetagem baseado no Minicorpus de treinamento, que há um salto final de acurácia, com 92,24% de acerto. O Quadro 2 a seguir ilustra a etiquetagem por classes de palavras com parte do trecho já apresentado no Quadro 2. Além da normalização ortográfica e etiquetagem por classes gramaticais, o Programa 1 fornece dados e plotagens automáticas já explicitadas de I a IX, no começo desta seção. É importante destacar que é o usuário que decide qual transcrição escolher, que pode ser individual ou do minicorpus inteiro.

Exemplos extraídos a partir do Programa 1

As palavras mais frequentes do MICONTES também podem ser visualizadas automaticamente (Figura 3). Nessa figura, são apagadas a maioria das palavras gramaticais e com menor sentido – chamadas de *stopwords* – e são preservadas as palavras mais lexicais. Isso significa que são plotados, basicamente, substantivos, verbos, adjetivos e advérbios. Se isso não for feito, ocorre o predomínio absoluto de palavras mais gramaticais (artigos, pronomes, preposições, etc) e as mais lexicais (substantivos, verbos, adjetivos, advérbios) simplesmente desapareceriam das visualizações por sua baixa frequência.[5]

Esse predomínio de *stopwords* é algo esperado em um *corpus*, tal como já foi verificado em outros trabalhos (COSTA JÚNIOR, 2022) para outros corpora orais do português, de forma que é algo constante que palavras de menor sentido sejam

[5] Caso o usuário queira eliminar as palavras gramaticais, pode utilizar a visualização sem *stopwords*, a qual elimina a maioria das palavras gramaticais das transcrições do Minicorpus.

mais abundantes e as de maior sentido sejam mais escassas. Entretanto, também fornecemos visualizações puras, sem nenhuma filtragem de *stopwords*, caso o usuário esteja interessado.

Ademais, é possível gerar visualizações de alguns fenômenos fonéticos e disfluências variadas da fala presentes nas transcrições, tais como truncamentos, alongamentos vocálicos e palavras com redução segmental, conforme já sinalizado no início da seção 3.5. Na Figura 4, por exemplo, é possível ver que a maioria das pausas ocorre diante de advérbios – a lista dessas palavras também é fornecida pelo programa – além de classes de gramaticais mais funcionais, tais como artigos, pronomes e conjunções coordenadas.

Quadro 2. Resultado da etiquetagem por classe de palavras

[('tinha', 'V'), ('um', 'ART'), ('homem', 'N'), ('que', 'PRO-KS-REL'), ('ele', 'PROPESS'), ('era', 'VAUX'), ('casado', 'PCP'), ('aí', 'ADV'), ('ele', 'PROPESS'), ('tinha', 'V'), ('um', 'ART'), ('filho', 'N'), ('que', 'PRO-KS-REL'), ('ele', 'PROPESS'), ('chamava', 'V'), ('Joãozinho', 'NPROP'), ('aí', 'ADV'), ('ele', 'PROPESS'), ('saiu', 'V'), ('para', 'PREP'), ('trabalhar', 'V'), ('o', 'ART'), ('o', 'ART'), ('velho', 'N'), ('deu', 'V'), ('abençoou', 'V'), ('ele', 'PROPESS')

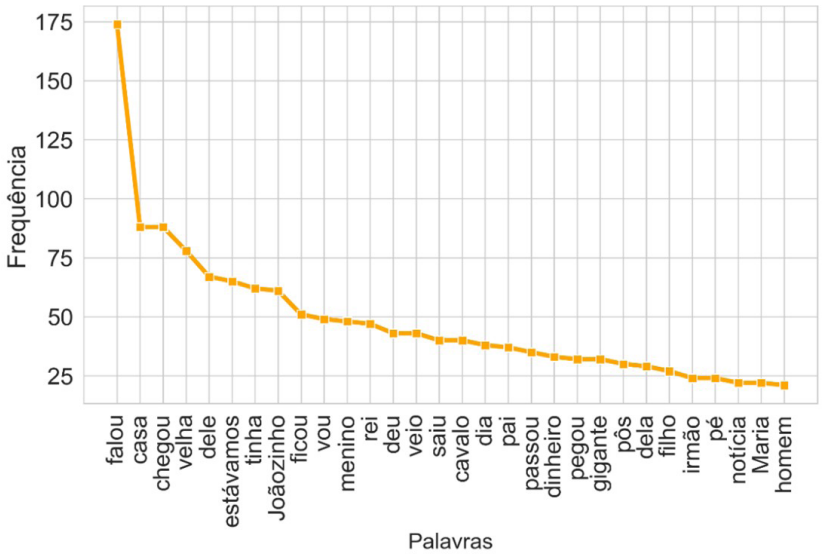

Figura 3. Palavras mais frequentes em MICONTES (sem *stopwords*)

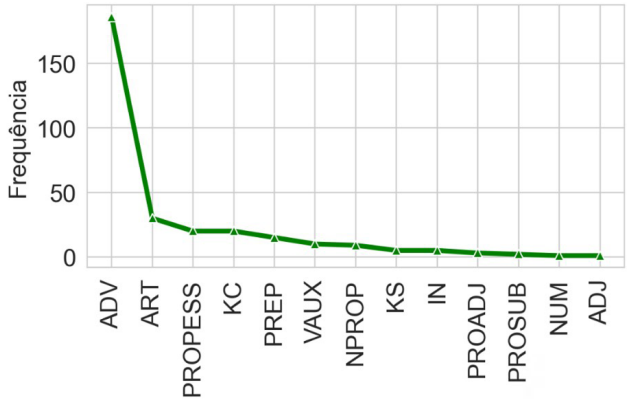

Figura 4. Classes de palavras em MICONTES

Programa para estudo do léxico

O programa para estudo do léxico e classes gramaticais oferece as funcionalidades a seguir.

I. Lematização (para classes variáveis) – com DELAF e Spacy.
II. Estemização da palavra com o RSLP Stemmer, da NLTK.
III. Detalhes morfológicos como flexões de verbos e nomes e classificação verbal bastante precisa a partir das informações do DELAF.
IV. Contagem da frequência de todos os tempos e modos verbais da transcrição em questão, assim como das flexões de substantivos e adjetivos;
V. Contagem de sílabas por meio da aplicação de regras fonológicas variadas inseridas no código com esta função.
VI. Classificação fonética do segmento inicial de classes gramaticais.
VII. Gráficos variados e automáticos, tais como nuvem de palavras, gráficos de linha com frequência de classes gramaticais, entre outros.

Para I, III e IV, foi criada uma conexão entre o DELAF e as transcrições do Minicorpus CE, que são utilizadas no Programa 2. Esse procedimento é descrito na seção a seguir.

Lematização e detalhes morfológicos com o Delaf

O DELAF utilizado[6] possui uma série de informações a respeito de palavras de classes lexicais, a saber, substantivos, adjetivos, verbos e advérbios. São 878.654 entradas que fornecem uma entrada, que pode ser flexionada, o lema dessa entrada e informações morfológicas, de acordo com sua classe gramatical. Isso pode ser visto no exemplo a seguir.

Exemplo 2:

digeriu,digerir.V:J3s

Nesse exemplo, a entrada é a forma verbal *digeriu*. A seguir, está o lema da palavra, que representa sua forma – verbal, neste caso – mais básica, que é o infinitivo. Depois, a classe à qual a entrada pertence, que é o Verbo (V). Posteriormente, há as informações do Tempo e Modo, que, neste caso, é o Pretérito Perfeito do Indicativo. A seguir, estão as flexões de pessoa e número, que são 3ª pessoa e singular, respectivamente.

A partir de operações de mineração de texto e limpeza de dados, foi possível separar as classes previamente etiquetadas e comparar sua forma com as do DELAF. Se são iguais, o Programa 2 vai requisitar as informações disponíveis do DELAF, de acordo com a classe gramatical em questão.

[6] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html

A partir dessas operações, é possível obter com extrema precisão a classe gramatical das palavras dos contos de encantamento, além de seu detalhamento morfológico. A seguir, exemplificamos com uma breve análise gerada a partir da contagem de substantivos e visualização de substantivos e verbos do minicorpus.

Exemplos de resultados obtidos pelo programa de análise do léxico

Em sua estrutura, os contos de encantamento apresentam uma sequência de fatos compostos por categorias de palavras. Os verbetes, pela voz do contador, demarcam quem são os personagens, tempo e espaço. Isso pode ser visto na Figura 5.

Como podemos observar na Figura 5, que representa os substantivos mais frequentes do minicorpus, a palavra *casa*, que apresenta maior recorrência, é um demarcador de lugar, assim como as palavras *palácio* e *serra*. A diferença entre elas está somente em relação ao contexto, em como aparecem na sequência do conto. A palavra *casa*, em geral, aparece logo no início dos contos e indicam um ambiente familiar. A *casa* é o espaço inicial em que se encontram o *pai*, *mãe*, *filhos* e *irmão*, não necessariamente obedecendo a uma composição comum em todos os contos, visto que há contos em que não tem a presença de irmão ou da figura materna. A *casa* será o ponto de partida do herói. Este herói está indicado pela segunda palavra, *menino*, além de *filho*, nas frequências subsequentes.

O substantivo *rei* é um indicativo de onde o herói deve chegar: ir até o rei, conquistar a princesa mais desejada. A *princesa***,** bem como os substantivos femininos correlatos, justificam, por sua vez, a repetição das palavras *mulher* e *moça*, que podem ser utilizados como sinônimos. A ida até o rei, que é o deslocamento do herói, implica na existência de um meio de locomoção, daí a constância da palavra *cavalo*. O cavalo

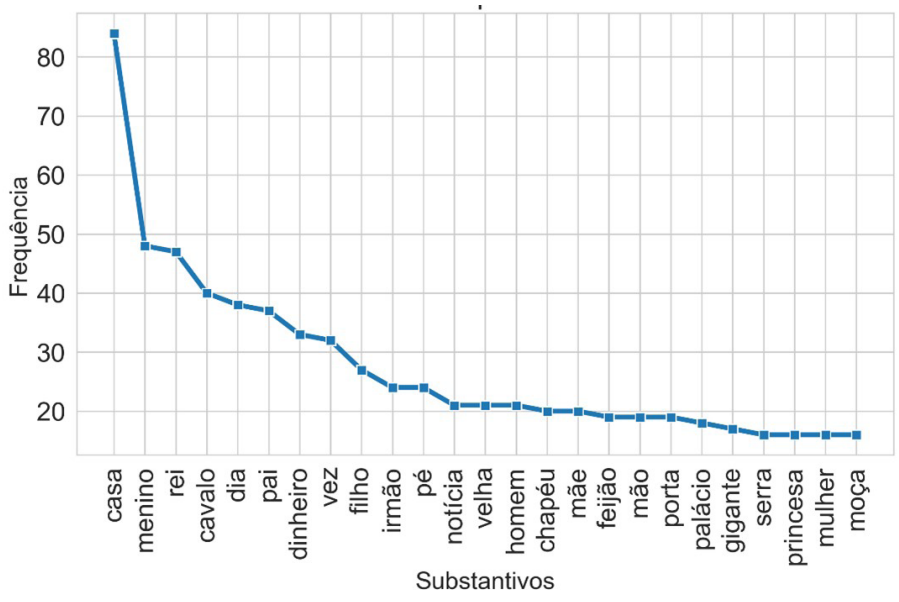

Figura 5. Substantivos mais frequentes em MICONTES

é um meio de locomoção próprio de um ambiente rural, interiorano, sendo talvez esse animal uma memória comum nos contos populares.

A palavra *velha* comumente indica uma personagem presente no meio do trajeto do herói. Pela afirmação de Cascudo (2006), a figura da velha personifica um vilão. Contudo, no minicorpus pesquisado, essa velha pode representar um personagem que apadrinha e ajuda o herói, em alguns contos sendo apontada pelo contador como Nossa Senhora (uma santa), uma figura divina, que irá socorrer o herói. Portanto, na casa da velha, uma senhora geralmente sábia e acolhedora, é onde o herói irá descansar e se alimentar. Essa velha, muitas vezes, é quem também fornecerá algum elemento mágico ao herói. Por fim, os verbetes *gigante* e *feijão*, indicam que se trata especificamente do conto "O pé de feijão". Essas palavras, respectivamente, abrangem o espaço do sobrenatural e do elemento mágico.

A Tabela 1 mostra características dos dez verbos mais frequentes do minicorpus. Essa tabela deve ser visualizada juntamente com a Figura 6, que mostra a classificação dessas formas verbais. Na tabela, além da frequência, é possível visualizar seu lema, o tamanho silábico e a raiz fornecida pela NLTK.

Paralelamente, é possível observar na Figura 6 que há grande predomínio de verbos no pretérito perfeito do indicativo (falou, chegou, veio), além do imperfeito (estávamos, tinha). Essa grande quantidade de verbos no pretérito perfeito e imperfeito pode refletir o predomínio da tipologia narrativa presente nos contos de encantamento, uma vez que o revezamento entre os dois aspectos gramaticais é um dos meios de mover a narrativa e inserir novos eventos. Além dos já referidos pretéritos, outra classificação verbal muito frequente foi de infinitivos, futuro do subjuntivo e gerúndios. A frequência dessas classificações também é bastante semelhante em *corpus* de fala do português, tal como discutido em Costa Júnior, 2022.

Tabela 1. Formas verbais mais frequentes em MICONTES

index	verbos	frequência	lema_nilc	syl_size	raiz_nltk
0	falou	169	falar	2	fal
1	chegou	84	chegar	2	cheg
2	estávamos	62	estar	4	est
3	tinha	47	ter	2	tinh
4	veio	43	vir	1	v
6	deu	42	dar	1	deu
7	saiu	40	sair	1	saiu
8	passou	32	passar	2	pass
9	pegou	31	pegar	2	peg

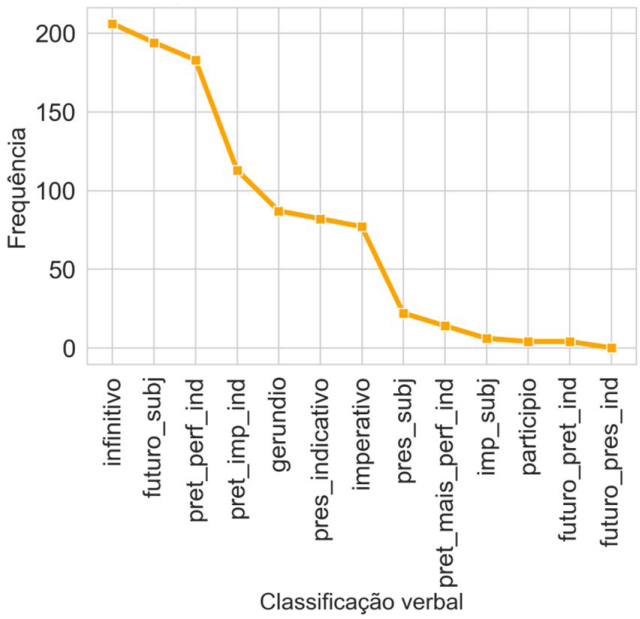

Figura 6. Classificação verbal mais frequente em MICONTES

Considerações finais

As narrativas orais podem ser consideradas um amplo campo de investigação dentro do meio acadêmico. Estudar a produção oral de grupos marginalizados é impulsionar para um movimento científico cada vez mais dinâmico e menos estigmatizado da língua falada – e contada – pelo interior do Brasil.

Desse modo, compreendemos que esta pesquisa contribui para reafirmar a existência do conto popular enquanto produção oral estreitamente vinculada à cultura de um determinado grupo social. O conto popular, ao ser um fenômeno linguístico-cultural, absorve as características da própria língua em uso: ganha novos contornos; adapta-se ao meio em que é produzido; absorve os objetos disponíveis no espaço físico e no espaço das experiências arquivadas pela mente; modela-se de acordo com a passagem do tempo; perde alguns adereços; ganha novas roupagens; mantém-se como memória viva e, constantemente, atualizada.

Nesse sentido, as tecnologias aqui disponibilizadas viabilizam sua exploração do ponto de vista linguístico e científico, de forma a obter padrões que mostram principalmente a diversidade lexical e gramatical que compõem os contos de encantamento.

Referências

ALUÍSIO, S.; PELIZZONI, J.; MARCHI, A.R.; (*et al.*). An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese. In: *Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language*. PROPOR, 2003.

BIRD, S.; LOPER, E.; KLEIN, E. *Natural Language Processing with Python.* O'Reilly Media Inc. Adaptado para Python 3 e disponibilizado gratuitamente, 2009.

CASCUDO, L. C. *Literatura oral no Brasil*. 2. ed. São Paulo: Global, 2006.

COSTA JÚNIOR, J. C. *Padrão informacional de stanzas de pacientes com esquizofrenia*. Tese de Doutorado. Faculdade de Letras, Programa de Pós Graduação em Linguística. Fevereiro de 2022. Disponível em: https://repositorio.ufmg.br/handle/1843/44004. Acesso em: 17 maio 2023.

FERREIRA, C. J. *Traz a lamparina e lumeia a cara do homem:* morfologia e construção de sentidos nas fábulas tocantinenses. Dissertação de Mestrado em Literatura e Práticas Sociais. Brasília: TEL/UnB, Departamento de Teoria Literária e Literaturas – UnB, 2010.

JUBRAN, C. C. S. A construção do texto falado. In: *Gramática do português culto falado no Brasil*. São Paulo: Contexto, 2015.

LIMA, N. C. *Narrativas orais:* uma poética da vida social. Brasília: Editora UnB, 2003.

MARCUSCHI, L. A. Domínios discursivos e gêneros textuais na oralidade e na escrita. In: *Produção textual, análise de gêneros e compreensão*. São Paulo: Parábola, 2008. p. 193-197.

MCKINNEY, W.; *et al*. Data structures for statistical computing in python. In: *Proceedings of the 9th Python in Science Conference.* 2010. p. 51–6.

MUNIZ, M. C. M. *Léxicos Computacionais:* Desafios na Construção de um Léxico de Português Brasileiro. Monografia de Qualificação. Instituto de Ciências Matemáticas de São Carlos, USP. 50p. Fev, 2003. Disponível em: http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/publicacoes.html. Acesso em: 17 maio 2023.

MUNIZ, M. C. M. *A construção de recursos lingüístico-computacionais para o português do Brasil:* o projeto de Unitex-PB. Dissertação de Mestrado. Instituto de Ciências Matemáticas de São Carlos, USP. 72p. 2004

PEREIRA, V. L. F. *O artesão da memória no Vale do Jequitinhonha*. Belo Horizonte: Editora UFMG; Editora PUC-Minas, 1996.

VAN ROSSUM, G.; DRAKE, J. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam; 1995.

VAN ROSSUM, G. *The Python Library Reference, release 3.8.2.* Python Software Foundation, 2020.

VIEIRA, V. P. S. *Narrativas orais do Alto Jequitinhonha*: uma proposta de análise tópica em contos populares do Serro. Dissertação (Mestrado Profissional – Programa de Pós-Graduação em Ciências Humanas) – Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, 2018. Disponível em: http://acervo.ufvjm.edu.br/jspui/handle/1/1892. Acesso em: 17 maio 2023.

WEITZEL, A. H. *Folclore literário e linguístico*: pesquisas de literatura oral e de linguagem popular. 2. ed. Juiz de Fora: EDUFJF, 1995.