DOI: https://doi.org/10.14393/DLv20a2026-21
e020021

Artigo

Palavras diacríticas e semidiacríticas em locuções adverbiais e adjetivas do português: um estudo a partir de *corpus*

Diacritic and semi-diacritic words in Portuguese adverbial and adjectival phrases: a corpus-based study

Luiz Leandro GOMES de LIMA[*]
Elizabete Aparecida MARQUES[**]

RESUMO: Este artigo investiga as palavras diacríticas e semidiacríticas do português contemporâneo (brasileiro e europeu), com foco em sua ocorrência em locuções adverbiais e adjetivas (à beça, à tona, de araque, à toa, com afinco, de soslaio, de araque, à paisana, do balacobaco, etc.). Essas palavras, caracterizadas por forte restrição fraseológica e por uma assemanticidade relativa, foram pouco investigadas nos estudos fraseológicos portugueses, embora sejam amplamente estudadas em outras tradições românicas. Partindo do referencial teórico da Fraseologia espanhola, o objetivo do trabalho é identificar de forma sistemática as lexias que funcionam como núcleos de locuções a partir de uma abordagem empírica e baseada em *corpus*. Para isso, foi compilado o Corpus do Português Sincrônico (CorPS), com cerca de 475 milhões de palavras provenientes de textos literários, jornalísticos, técnico-científicos e cotidianos escritos no século XXI. Utilizou-se um Índice de Restrição Fraseológica (IRF), calculado automaticamente por meio de um *script* em Python e revisado por um crivo manual. O IRF determina a proporção de ocorrências de cada palavra antecedidas por preposição em relação ao total de ocorrências no *corpus*. Foram consideradas diacríticas as palavras com IRF de 100% e semidiacríticas aquelas com IRF igual ou superior a 80%, desde que registradas em dicionários gerais da língua e presentes em mais de uma formação discursiva. Foram identificadas 28 palavras diacríticas e 109 semidiacríticas, confirmando que a restrição fraseológica total é rara e que há um *continuum* de fossilização lexical no português contemporâneo. A pesquisa evidencia o potencial das ferramentas da Linguística de Corpus para delimitar empiricamente fenômenos tradicionalmente descritos de modo intuitivo, demonstrando que a integração entre critérios estatísticos e lexicográficos permite propor um modelo replicável para o estudo da fixação lexical. Conclui-se que as palavras diacríticas, embora representem um fenômeno marginal na fraseologia portuguesa, desempenham papel relevante na compreensão dos processos de cristalização e mudança do léxico.
PALAVRAS-CHAVE: Fraseologia. Palavras diacríticas. Léxico. Linguística de Corpus. Português contemporâneo.

[*] Doutorando em Estudos de Linguagens pela Universidade Federal de Mato Grosso do Sul (UFMS), Campo Grande, Mato Grosso do Sul – Brasil. luiz.gomes@ufms.br
[**] Doutora em Linguística Aplicada pela Universidad de Alcalá de Henares. Professora Titular Emérita da Universidade Federal de Mato Grosso do Sul (UFMS), Campo Grande, Mato Grosso do Sul – Brasil. eamarques@hotmail.com

ABSTRACT: This article deals with diacritic and semi-diacritic words in contemporary Portuguese (both Brazilian and European), focusing on their occurrence in adverbial and adjectival phrases (e.g., à beça, à tona, de araque, à toa, com afinco, de soslaio, à paisana, do balacobaco, among others). These words, characterized by strong phraseological restriction and relative asemanticity, have been little explored in Portuguese phraseological studies, although they are widely studied in other Romance traditions. Drawing on the theoretical framework of Spanish Phraseology, this study aims to systematically identify lexical items that function as the nuclei of set phrases through an empirical, corpus-based approach. For this purpose, the Corpus of Contemporary Portuguese (CorPS) was compiled, comprising about 475 million words from literary, journalistic, scientific-technical, and everyday texts written in the 21st century. A Phraseological Restriction Index (IRF) was automatically calculated using a Python script and subsequently reviewed manually. The index measures the proportion of occurrences of each word preceded by a preposition relative to its total occurrences in the corpus. Words with a PRI of 100% were classified as diacritic, and those with a PRI equal to or greater than 80% as semi-diacritic, provided they are recorded in general dictionaries of the language and occur in more than one discourse formation. A total of 28 diacritic and 109 semi-diacritic words were identified, confirming that full phraseological restriction is rare and that a continuum of lexical fossilization exists in contemporary Portuguese. The study highlights the potential of Corpus Linguistics tools to empirically delimit phenomena traditionally described intuitively, showing that the integration of statistical and lexicographic criteria allows the proposal of a replicable model for studying lexical fixation. We conclude that diacritic words, although representing a marginal phenomenon within Portuguese phraseology, play a relevant role in understanding processes of lexical crystallization and change.
KEYWORDS: Phraseology. Diacritic words. Lexicon. Corpus Linguistics. Contemporary Portuguese.

Artigo recebido em: 19.11.2025
Artigo aprovado em: 02.05.2026

1 Introdução

Os estudos fraseológicos se ocupam de explicar as Unidades Fraseológicas (Corpas Pastor, 1996) e lançam luz sobre seu papel central no funcionamento das línguas naturais. As Unidades Fraseológicas (UFs) são definidas como "Construções pluriverbais forjadas diacronicamente a partir daquelas combinações de palavras que acabaram por se tornar gradualmente estáveis em seu uso e que apresentam um nível variável de idiomaticidade com resultado metaforizado, bem como uma transparência ou opacidade que oscila conforme os casos" (Echenique Elizondo, 2021, p. 44, tradução

nossa)[1]. Dentre essas unidades, podemos destacar aquelas que têm como núcleo palavras de uso restrito, que só ocorrem em contextos de grande fixidez e que, em certa medida, não apresentam significação lexical autônoma. Conhecidas como palavras diacríticas (Zuluaga, 1980; Ruiz Gurillo, 1998)[2], essas lexias não circulam livremente pelo sistema linguístico, sendo encontradas exclusivamente ou quase exclusivamente, na sincronia atual da língua, em unidades fraseológicas como colocações (*redondamente enganado*, *terminantemente proibido*), locuções (*à beça*, *à tona*, *de araque*) ou enunciados fraseológicos (*quem não arrisca, não petisca*)[3]. Sua principal característica é a restrição fraseológica, que frequentemente vem acompanhada de uma assemanticidade relativa, o que as torna difíceis de tratar dentro das categorias tradicionais da morfologia e da semântica lexical.

[1] Construcciones pluriverbales diacrónicamente acuñadas a partir de aquellas combinaciones de palabras que han terminado por ser gradualmente estables en su uso y presentar un nivel variable de idiomaticidad con resultado metaforizado, así como una transparencia u opacidad que fluctúa según los casos.

[2] Essas palavras já foram designadas por outros termos diversos: na fraseologia espanhola, também são conhecidas por palavras idiomáticas (Garcia-Page, 1990, 1991; Zuluaga, 1992), *hapax* fraseológicos (González Rey, 2005) e palavras presas (Echenique Elizondo, 2021); na tradição inglesa, são conhecidas como *cranberry words* (Aronoff, 1976); na tradição alemã, por *unikale Elemente* (Holzinger, 2012); na tcheca, como palavras monocolocáveis (Čermák et al, 2016). O termo "diacrítico" foi proposto por Zuluaga (1980): se a palavra diacrítica só pode existir dentro de uma UF, o aparecimento dessa palavra implica que o sintagma em questão é uma UF, ou seja, ela é um índice inequívoco (diacrítico) de que estamos diante de uma UF.

[3] Adotamos neste trabalho a classificação de UFs proposta por Corpas Pastor (1996). A autora identifica cinco critérios básicos presentes em classificações fraseológicas anteriores: (1) ser elemento oracional ou oração completa; (2) apresentar fixação no sistema, na norma ou na fala; (3) ser fragmento de enunciado ou enunciado completo; (4) ter restrição combinatória limitada ou total; e (5) o grau de motivação semântica. Considerando que nenhum desses critérios serve isoladamente para estruturar uma taxonomia global, Corpas Pastor (1996) seleciona e combina dois deles: a capacidade de constituir um enunciado (e, consequentemente, um ato de fala) e o tipo de fixação (na língua, na norma, na fala). A partir do cruzamento desses dois parâmetros, ela organiza o universo fraseológico em três esferas: colocações (não constituem enunciados, estão fixadas na norma, são sintagmas livres com baixo grau de restrição combinatória, são parcialmente composicionais), locuções (não constituem enunciados, estão fixadas no sistema da língua, têm alto grau de coesão semântica e seu sentido não se justifica pela soma dos significados de seus componentes) e enunciados fraseológicos (constituem enunciados completos e atos de fala por si mesmos, estão fixados na fala, são unidades do acervo sociocultural da comunidade de fala, podem se dividir em parêmias e fórmulas rotineiras).

Apesar da relevância linguística das palavras diacríticas, sobretudo pelo que revelam sobre a história da língua, quase não há considerações científicas sobre elas em português. Além disso, os poucos estudos se baseiam em listas compiladas a partir de observações intuitivas, e não em levantamentos empíricos da ocorrência e distribuição desses itens lexicais. Neste artigo, propomos uma metodologia baseada em *corpus* com o objetivo de identificar semiautomaticamente palavras diacríticas e semidiacríticas em unidades fraseológicas do português. Nosso interesse se centra nas palavras diacríticas presentes em locuções adverbiais (*à toa*, *com afinco*, *de soslaio*) e adjetivas (*de araque*, *à paisana*, *do balacobaco*) do português contemporâneo.

Na primeira seção, apresentamos os pressupostos teóricos que fundamentam esta pesquisa, destacando a relevância do estudo das palavras diacríticas no âmbito da Fraseologia Histórica. A segunda seção é dedicada à descrição dos procedimentos metodológicos adotados. Na terceira, expomos os resultados obtidos e promovemos uma breve discussão a partir deles. Por fim, na quarta seção, apresentamos as considerações finais.

2 Palavras diacríticas: um problema de Fraseologia Histórica

A base teórica utilizada neste artigo advém da escola de Fraseologia espanhola, que tem uma ampla tradição de estudo das palavras diacríticas (cf. Zuluaga, 1980; Garcia-Page, 1990, 1991, 2008; Aguilar Ruiz, 2011, 2020; Lorenzo, 2021). Já em língua portuguesa, não há estudos sistemáticos sobre essas palavras, tendo sido apenas mencionadas em trabalhos com outros objetos:

- Biderman (2001, p. 47), ao tecer considerações sobre o dicionário Aurélio, denomina de "fósseis lexicais" palavras como *guisa* e *soslaio* e critica a obra por ter entradas independentes para esses vocábulos, "já desaparecidos de há muito da língua portuguesa".

- Borba (2003) refere-se a essas palavras como "nomes presos" ao explicar as diferentes formas sintáticas dos sintagmas fixos que podem figurar em obras lexicográficas e traz uma pequena lista de palavras diacríticas: *chofre*, *roldão*, *cor* (*de cor*), *supetão*, *cambulhada*, *socapa*, *ínterim*, *triz*.
- Fulgêncio (2008, p. 114-115) menciona brevemente o fenômeno ao relacionar as idiossincrasias lexicais das unidades fraseológicas e apresenta uma lista de 58 palavras diacríticas, embora dentre elas haja componentes de palavras compostas (*bel-prazer*, *mal-humorado*) e outras palavras que não podem ser consideradas diacríticas (como *crivo*, que a autora identifica como pertencente exclusivamente à locução *passar pelo crivo de*, mas que, em verdade, tem uso autônomo na língua).
- Simão (2014) e Caniato e Simão (2015) discutem a lematização das palavras diacríticas em dicionários bilíngues português/ espanhol e as soluções tradutológicas para as unidades fraseológicas com essas palavras.

O primeiro autor a tratar de maneira sistemática as palavras diacríticas em espanhol foi Zuluaga (1980). Buscando demonstrar que as unidades fraseológicas, em geral, não se conformam às regras do "discurso livre", o autor argumenta que essas unidades podem apresentar anomalias de ordem sintática (como no caso da locução *chorar pitangas*, que infringe a regência habitual do verbo *chorar*), bem como de ordem lexical, ao incluírem termos inexistentes no léxico geral da língua. Esses termos, denominadas por ele de palavras diacríticas, seriam "carentes de toda autonomia semântica, reconhecidas pelo falante somente dentro de expressões fixas (*lirondo, contera, vilo*)" (Zuluaga, 1980, p. 102, tradução nossa)[4].

[4] Carentes de toda autonomía semántica, reconocidas por el hablante solamente dentro de expresiones fijas (lirondo, contera, vilo)". *Lirondo*, palavra proveniente do latim *limpidus*, compõe em espanhol a UF *mondo y lirondo*, que significa "limpo", "sem acréscimos"; *contera*, palavra derivada de *conto*, compõe em espanhol UFs como por exemplo *echar la contera*, que significa "concluir um negócio ou discurso"; *vilo*, palavra de origem incerta, compõe em espanhol a UF *en vilo* que significa "bamboleante", "sem o suficiente apoio (físico).

Zuluaga (1980) estabelece, assim, as duas propriedades características das palavras diacríticas: assemanticidade e restrição fraseológica. Por assemanticidade entendemos a falta de significado léxico das palavras diacríticas: para a maior parte dessas lexias, é impossível lhes atribuir um significado (Garcia-Page, 1990, p. 280). Nesse sentido, podem ser descritas como signos fonológicos sem significado lexical pleno; somente a UF, em sua totalidade, tem um significado que, como é comum nos fraseologismos, não deriva da soma dos significados dos seus itens constitutivos. Uma forma de demonstrar essa assemanticidade é a dificuldade em "desautomatizar" as unidades fraseológicas com palavras diacríticas: enquanto é totalmente possível imaginar um cenário literal para a locução *bater as botas* ("colidir os calçados"), não é possível fazer o mesmo para *ir para o beleléu*, pois *beleléu* é uma palavra diacrítica.

Importa observar, contudo, que a assemanticidade dessas palavras não é absoluta. Trata-se de um *continuum*, onde certas formas como *beça*, *tona* ou *araque* se situam num extremo opaco, enquanto outras, como *contento*, *repente*, ou *antemão* ainda evocam imagens conceituais vagamente associadas à forma sonora. É perfeitamente possível reconhecer em *matina* a mesma raiz de *matutino*, ou relacionar *repente* com *repentino*, ou ainda identificar o falso prefixo *ante* em *antemão* (Garcia-Page, 1990, chama essas palavras de "elementos virtuais", pois são compreensíveis ao falante isoladamente). Uma vez que a assemanticidade pode ser relativizada, a restrição fraseológica se torna o critério central na definição das palavras diacríticas, sendo justificável seu estudo dentro da Fraseologia[5].

Do ponto de vista diacrônico, no entanto, a restrição fraseológica também pode ser pensada em termos de gradação. Uma das principais fontes de palavras diacríticas é a história da língua: são as chamadas "palavras fósseis", arcaísmos vindos de fases pretéritas da língua e que sobreviveram até a sincronia atual incrustadas em uma

[5] Disciplina que se interessa pelas unidades lexicais que (i) constam de ao menos duas palavras gráficas, (ii) apresentam certo grau de lexicalização e (iii) se caracterizam pela alta frequência de coaparição na língua (Corpas Pastor, 1996, p. 18).

Unidade Fraseológica (Garcia-Page 1990, p. 285; Aguilar Ruiz, 2011, 2020; Lorenzo, 2021). É válido supor (e é possível demonstrar por meio de *corpora* históricos) que em sincronias anteriores essas palavras tenham tido autonomia sintática e plenitude semântica, e que essas características foram se perdendo na evolução da língua, ao mesmo tempo em que a distribuição da palavra foi ficando mais restrita[6]. Sincronicamente, portanto, é possível falar em palavras semidiacríticas, que são aquelas que mantêm um número reduzido de usos fora da locução em que geralmente aparecem. Elas podem ser vistas como formas em vias de fossilização, ou seja, em processo de se tornarem plenamente diacríticas. Garcia-Page (1990) chama esses termos de "signos de transição ao arcaísmo", por estarem em um estado intermediário entre o uso autônomo e a restrição típica das palavras plenamente diacríticas. Um exemplo em português é *afinco*, que aparece em mais de 95% das ocorrências na locução *com afinco*, mas ainda é reconhecível como substantivo autônomo por parte dos usuários da língua[7].

Tendo em vista que a perspectiva diacrônica é essencial para entender o processo de formação das unidades fraseológicas com palavras diacríticas, é necessário recorrer aos conceitos e pressupostos teóricos da Fraseologia Histórica, campo responsável por rastrear o percurso diacrônico das expressões cristalizadas da língua (Echenique Elizondo, 2003, 2021, 2023). Apenas uma abordagem que leve em conta a dimensão temporal e os mecanismos de fossilização lexical pode explicar adequadamente a emergência e o comportamento das palavras diacríticas. Neste

[6] A título de exemplo podemos citar a palavra *prol*, que, conforme veremos em 4.1, é uma palavra totalmente restrita fraseologicamente na sincronia atual do português, mas que em sincronias passadas gozava de autonomia lexical, significando "ganho" ou "benefício": "E mia senhor, se eu morrer', dizede-me ¿que *prol* vus á?", "Per bõa fé non é meu ben, nen é mha *prol* viver assy", "Diz' ela: 'Nom vos tem *prol* esso que dizedes'" (Cantigas de Amigo, século XIII, Corpus Informatizado do Português Medieval, disponível on-line em: https://cipm.fcsh.unl.pt/, acesso em: 13 out. 2025).

[7] Alguns exemplos de uso autônomo retirados do CorPS (veja 3.1): "Corey voltou a beijá-la no mesmo *afinco* de antes" (discurso literário), "já eles, apesar do *afinco*, provavelmente não sonhavam em trabalhar com oque trabalham" (discurso cotidiano), "Aqui, casas como Chef Vivi se notabilizaram pelo *afinco* dedicado ao tema" (discurso jornalístico), "Esse *afinco* do professor em suas práticas cotidianas é um dos pontos enfatizados pelo Prêmio" (discurso técnico-científico).

trabalho temos como objetivo apenas listar e analisar sincronicamente as palavras diacríticas mais comuns do português, e por isso, aqui, não fazemos menção a seu devir histórico. No nosso caso específico, estamos interessados nas palavras diacríticas enclausuradas em locuções adverbiais e adjetivas do português (brasileiro e europeu) contemporâneo. Essas locuções apresentam, de modo geral, uma estrutura sintática relativamente fixa: consistem em um sintagma preposicional que seleciona um sintagma nominal, sendo este último tipicamente formado por um núcleo (a palavra diacrítica) e, ocasionalmente, por modificadores. Assim, a estrutura formal mais comum é [Prep [SN]], como em *à tona*, *de soslaio*, *com afinco*, *à paisana*, entre outras. As preposições mais frequentemente associadas a esse tipo de construção são "a", "de", e "em"[8], refletindo uma tendência que também se observa no espanhol (cf. Corpas Pastor, 1996, p. 99; Garcia-Page, 2008, p. 115-129). Em muitos casos, uma mesma palavra diacrítica pode ser selecionada por mais de uma preposição, como ocorre com *cima*, que aparece em locuções como *em cima*, *por cima*, *de cima*. Embora algumas locuções admitam modificadores internos (*com muito afinco*), outras não permitem variações, sendo altamente resistentes à inserção de elementos no seu interior (**à grande toa*, **à maior beça*), o que indica um grau mais alto de cristalização formal e semântica.

3 Procedimentos metodológicos

A maioria dos trabalhos sobre palavras diacríticas parte de compilações baseada em dicionários ou em listas de caráter intuitivo elaboradas por outros autores (cf. por exemplo Garcia Page, 1991, para o espanhol ou Dobrovorskij e Piirainen, 1994, para o alemão). No entanto, acreditamos que, para avançar na delimitação e no entendimento desse fenômeno, é necessária uma abordagem empírica baseada nos

[8] Conforme levantamento feito a partir das locuções adverbiais e adjetivas codificadas em Borba (2002): 67% das locuções adjetivas são encabeçadas pela preposição *de*, 17% pela preposição *em* e 10% pela preposição *a*; entre as adverbiais, 30% têm a preposição *a*, 29%, *em*, e 23%, *de*.

preceitos e metodologias elaborados pela Linguística de Corpus, que permitam identificar palavras diacríticas de forma semiautomática, a partir de critérios quantificáveis. Por isso, para este estudo, compilamos o Corpus do Português Sincrônico (CorPS), com 475 milhões de palavras e constituído de amostras da língua portuguesa escrita no século XXI. Nas próximas seções detalhamos o processo de elaboração do *corpus* e sua análise em busca de palavras diacríticas e semidiacríticas.

3.1. Compilação do Corpus do Português Sincrônico (CorPS)

Adotamos como referência teórica para planejamento e compilação do CorPS o modelo criado por Egbert, Biber e Gray (2022). De acordo com os autores, um *corpus* é definido como "uma grande e criteriosa amostra de textos projetado para representar um domínio do uso linguístico (e.g.: uma língua, um dialeto, um registro)" (Egbert; Biber; Gray, 2022, p. 7, tradução nossa)[9]. A principal característica de um *corpus* é sua representatividade, isto é, "a medida em que um corpus permite generalizações precisas sobre os padrões linguísticos quantitativos típicos de uma língua-alvo ou domínio discursivo" (Egbert; Biber; Gray, 2022, p. 11, tradução nossa)[10]. Dessa forma, o *corpus* deve ser projetado tendo em vista fenômenos da linguagem bem delimitados e um problema de pesquisa específico, o que implica fazer escolhas conscientes sobre os gêneros, registros e domínios a serem incluídos, de modo a garantir a representatividade dos fenômenos em questão. O nosso *corpus* foi construído com o objetivo principal de permitir a identificação de palavras diacríticas, embora possa ser útil também para estudos fraseológicos mais amplos e outras investigações linguísticas.

As primeiras considerações a serem feitas se referem ao domínio que o *corpus* deve representar. No nosso caso, o domínio é a língua portuguesa: utilizada por 260

[9] a large and principled sample of texts designed to represent a target domain of language use (e.g., a language, dialect, or register).

[10] the extent to which a corpus permits accurate generalizations about the quantitative linguistic patterns that are typical in a target language or discourse domain.

milhões de pessoas nos cinco continentes, sendo língua oficial em nove países (Camões - Instituto da Cooperação e da Língua, 2023). Todos esses usuários produzem diariamente uma infinidade de textos falados e escritos em diversos gêneros discursivos. Não é possível afirmar quando a língua portuguesa surgiu, mas os primeiros registros escritos datam do século XIII (Teyssier, 2001).

Como não é possível incluir no *corpus* toda a diversidade de textos já produzidos em língua portuguesa, o próximo passo é operacionalizar o domínio, isto é, identificar os estratos da língua para os quais é possível obter textos a serem coletados e incluídos na amostra, tentando ao máximo representar a variedade linguística do domínio. Optamos por incluir apenas textos já em formato digital, em sua maioria disponíveis gratuitamente na Internet. De todos os países lusófonos, optamos por incluir textos apenas do Brasil e de Portugal, por dois motivos: (i) esses são os únicos países em que a maioria da população é usuária nativa da língua portuguesa (Bagno, 2014, pág. 43); e (ii) a maior parte dos textos disponíveis *online* é de um desses dois países, sendo a grande maioria em língua portuguesa do Brasil (por isso também definimos como objetivo que 70% do *corpus* seria composto por textos do português brasileiro e 30% do português europeu).

Em relação aos gêneros que seriam incluídos, adaptamos a concepção elaborada por Costa (2008), que agrupa os gêneros em formações discursivas. Para representar nosso domínio, selecionamos as formações literária (romances, contos, poemas etc.), cotidiana (conversação e seus tipos), jornalística (notícias, reportagens, entrevistas etc.) e técnico-científica (artigos, resenhas, manuais etc.). Para facilitar o trabalho de coleta dos textos, utilizamos sempre que possível *datasets* já prontos, isto é, conjuntos de textos já raspados da Internet e disponibilizados em repositórios públicos. Para a formação discursiva *jornalística*, utilizamos conjuntos de dados extraídos de portais

jornalísticos como Folha de São Paulo[11], BBC Brasil[12] e de vários portais jornalísticos portugueses[13], disponibilizados em vários repositórios públicos. Para a formação discursiva *cotidiana*, utilizamos *datasets* de interações em comunidades de língua portuguesa do site Reddit, disponibilizadas pelo repositório Academic Torrents[14]. Para as formações discursivas *literária* e *técnico-científica*, não encontramos conjuntos de textos já prontos, e por isso tivemos de utilizar técnicas de raspagem para extrair textos dos *sites* Wattpad[15] e Scielo[16]. Em todos os casos, controlamos rigorosamente a origem dos textos, de modo a separar aqueles produzidos em português brasileiro dos escritos em português europeu. No que se refere ao período de produção, selecionamos apenas textos escritos no século XXI.

O CorPS em sua forma final contém cerca de 475 milhões de palavras, com uma proporção de 25% para cada formação discursiva. Esse tamanho de *corpus* é suficiente para generalizar estimativas de uso para grande parte das palavras diacríticas atestadas em circulação.

---

[11] *Dataset* "Folha - News of the Brazilian Newspaper - 2024", que contém artigos do portal da Folha de São Paulo publicados entre 2017 e 2024, disponibilizado pelo usuário luisfcaldeira no repositório Kaggle (disponível em: https://www.kaggle.com/datasets/luisfcaldeira/folha-news-of-the-brazilian-newspaper-2024, acesso em: 11 abr. 2025), totalizando 76,7 milhões de palavras.

[12] *Dataset* "celsowm/bbc_news_ptbr", que contém artigos do portal BBC Brasil, disponibilizado pelo usuário Celso F. no repositório Hugging Face (disponível em: https://huggingface.co/datasets/celsowm/bbc_news_ptbr, acesso em: 29 jun. 2025), totalizando 7,7 milhões palavras.

[13] *Dataset* "FEUP news corpus", que contém artigos de portais portugueses, disponibilizado pelo pesquisador Henrique Lopes Cardoso no repositório Portulan Clarin (disponível em: https://portulanclarin.net/repository/browse/feup-news-corpus/a62ead58a94011e9a62602420a000003475e75ce3d3a4880b4d58cd9c56e91da/, acesso em: 17 jun. 2025), totalizando 37,9 milhões de palavras.

[14] Disponível em https://academictorrents.com/details/ba051999301b109eab37d16f027b3f49ade2de13, acesso em: 11 abr. 2025. Foram incluídas todas as interações entre os usuários das comunidades "r/conversas" e "r/PergunteReddit" para o português brasileiro (totalizando 88,8 milhões de palavras) e "r/CasualPT" para o português europeu (36,9 milhões de palavras).

[15] Portal onde os usuários compartilham histórias ficcionais de autoria própria (disponível em https://www.wattpad.com/). Total: 87 milhões de palavras. A extração foi feita entre 12 maio 2025 e 16 maio 2025. Além das histórias do Wattpad, incluímos no *corpus* obras da literatura brasileira e portuguesa contemporânea (24,3 milhões de palavras).

[16] Indexador de revistas científicas (disponível em https://www.scielo.br/). Total: 114,6 milhões de palavras. A extração foi feita em 22 abr. 2025.

3.2. Análise do *corpus*

Para encontrar as palavras diacríticas no CorPS, utilizamos um *script*[17] de Python que percorre os textos do *corpus* e calcula o Índice de Restrição Fraseológica (IRF) das palavras, isto é, a porcentagem de ocorrências da palavra antecedida de preposições que estejam até três posições antes da palavra sob suspeita de ser diacrítica em relação ao total de ocorrências da palavra. Por exemplo: a palavra *repente* aparece 27.786 vezes no *corpus*. Em 27.696 dessas ocorrências (99,67%), essa palavra aparece antecedida da preposição *de* na locução adverbial *de repente*. As outras ocorrências estão assim divididas:

- 65 ocorrências (0,23%) antecedidas da preposição *em* (*num repente*, *em um repente*);
- 11 ocorrências (0,039%) de *repente* como substantivo que significa "ânsia" ou "ímpeto" ("Então me bateu um *repente* de que meus pais iriam morrer um dia");
- 7 ocorrências (0,025%) como o nome do estilo musical; e
- 7 ocorrências (0,025%) como um advérbio isolado ("Mas tudo foi gradativo, não foi *repente* que tive ânimo pra manter essa rotina"; "merda ver um filme que estas mesmo a gostar e *repente* corta para o intervalo").

Dessa forma, a palavra *repente* tem um IRF de 99,67%, pois, das 27.786 vezes que ocorre no *corpus* em análise, em 27.696 está antecedida de uma preposição

[17] A análise estatística e o cálculo do Índice de Restrição Fraseológica (IRF) foram realizados por meio de um *script* customizado em linguagem Python 3.10. O script utiliza as bibliotecas nativas *re* (expressões regulares) para limpeza de tags XML/HTML e tokenização, e *collections* para processamento de frequências. O cálculo de IRF é feito via janela retrospectiva: para cada *token* alvo, o algoritmo varre uma janela de adjacência de até três posições à esquerda, interrompendo a busca imediatamente após a identificação da primeira preposição (conforme léxico de 44 preposições e contrações do português: a, ante, após, até, com, contra, de, desde, em, entre, para, per, perante, por, sem, sob, sobre, trás; do, da, dos, das, deste, desta, daquele, daquela, no, na, nos, nas, num, numa, neste, nesta, naquele, naquela, pelo, pela, pelos, pelas, ao, aos, à, às.). Depois disso, aplicam-se filtros de significância: são retidas apenas palavras com frequência absoluta superior a 10 ocorrências no *corpus* e IRF superior a 0,80 (80%). O código-fonte permite a extração automatizada de metadados geográficos e de gênero textual baseados na nomenclatura dos arquivos do *corpus*.

formando uma locução adverbial (embora os casos de *num repente* e *em um repente* também envolvam antecedentes preposicionais, essas locuções não estão codificadas em nenhum dicionário consultado). Como nosso objetivo é buscar por palavras diacríticas que sejam núcleo de locuções adverbiais e adjetivas, essa abordagem simples que busca por preposições antecedentes é suficiente. Caso quiséssemos buscar por todas as palavras diacríticas, em qualquer contexto sintático, uma abordagem mais robusta seria necessária (como a utilizada por Čermák e Obstová, 2021, para buscar palavras diacríticas no espanhol e italiano).

Definimos que, para ser considerada uma palavra diacrítica, a lexia deve ter um IRF de 100%. Para as palavras semidiacríticas, definimos arbitrariamente uma proporção de pelo menos 80%, seguindo Holzinger (2018). Por isso, definimos que o *script* deveria retornar apenas palavras com 80% ou mais de IRF. Além disso, consideramos que a palavra só é considerada diacrítica ou semidiacrítica se for utilizada na língua geral, não tendo uso restrito a uma formação discursiva. Sendo assim, definimos também que o *script* deveria retornar apenas palavras que aparecem em textos de pelo menos duas das quatro formações discursivas que compõem o *corpus* e em pelo menos 5 (cinco) dos seus 1300 arquivos de texto. Por fim, como uma das preocupações da Linguística de Corpus é ter uma quantidade suficiente de dados para tornar possível a generalização a partir da amostra, consideramos apenas as palavras que ocorram com uma frequência total de 10 (dez) vezes ou mais no CorPS.

Após a análise automática e a extração das palavras candidatas a diacrítica/ semidiacrítica, realizamos ainda um crivo manual com o objetivo de corrigir erros decorrentes de ambiguidades ou inconsistências do *corpus*. Foi necessário excluir manualmente alguns casos atípicos, como:

- latinismos e outros estrangeirismos como *priori, posteriori*, *bono*, *cappella* e *carte*, que aparecem em locuções completas emprestadas do latim *a priori*, *a posteriori* e *pro bono*, do italiano *a cappella* e do francês *à la carte* (diferentemente de *ínterim*,

por exemplo, que, apesar de ser uma palavra emprestada do latim, ocorre em uma locução de base portuguesa: *nesse ínterim*);

- palavras enganosamente detectadas como diacríticas por aparecerem com frequência após o artigo feminino *a* ou o pronome átono *nos* (que está fortemente associada a verbos pronominais), confundidos pelo *script* com a preposição *a* e com a contração *em* + *os* = *nos* (como *(a) maioria* ou *(nos) preocupamos*);
- nomes próprios que ocorrem com frequência após preposições, como nomes de meses, certos sobrenomes (*(de) Saussure*, *(da) Vinci*, *(de) Arimatéia*) ou nomes que fazem parte de locuções (*(calcanhar de) aquiles*, *(lei de) talião*);
- erros ou adaptações ortográficas, como *favooor*, que aparece 20 vezes no *corpus*, todas antecedidas da preposição *por* (trata-se de uma tentativa de indicar na escrita o alongamento da sílaba tônica ao utilizar a locução *por favor* com uma entonação mais longa e enfática);
- palavras que, pela chance ou por sua distribuição, apareceram no *corpus* sempre depois de preposições, mas sem formar locuções fixadas, como pronomes oblíquos (*(de/ em) mim*, *(de/ em) si*). Um grupo grande dessas palavras inclui as que indicam materiais de fabricação, ingrediente ou origem geográfica (*(de) flandres*, *(de) vime*, *(de) cetim*, *(de) lycra*, *(de) veludo*, *(de) camomila*, *(de) copaíba*, *(da) Tasmânia*, *(do) Atacama*, *(de) Estocolmo*). Adotamos a institucionalização da unidade como critério de delimitação (Corpas Pastor, 1996; Garcia-Page, 2008): caso a unidade encontrada esteja codificada em um dicionário geral do português, ela é incluída como locução. Consultamos as seguintes obras: Dicionário Houaiss da Língua Portuguesa (Houaiss, *online*), Dicionário Caldas Aulete (Aulete Digital, *online*), Dicionário da Academia de Ciências de Lisboa (Academia das Ciências de Lisboa, *online*), Novo Dicionário Aurélio da Língua Portuguesa (Ferreira, 2010), Dicionário de Usos do Português do Brasil (Borba, 2002) e Dicionário da Língua Portuguesa da Porto Editora (Porto Editora, 2015);

- elementos de composição unidos por hífen, que o *script* encara como palavras separadas (*nobis* em *ora-pro-nobis*, sempre depois de *pro*);
- palavras que são diacríticas, mas na verdade fazem parte de uma locução nominal (*nabal* em *sol na eira e chuva no nabal*, *butchaca* em *tchaca tchaca na butchaca*) ou verbal (*veneta* em *dar na veneta, jus* em *fazer jus a*). Tendo em vista a dificuldade em separar locuções adverbiais de locuções verbais que têm uma locução adverbial dentro de si (Garcia-Page, 2008, p. 128-129), adotamos também o critério da institucionalização: se a locução aparece como adverbial ou adjetiva nos dicionários consultados, sua palavra diacrítica é incluída na lista; caso contrário, não.

Além disso, a análise manual dos resultados foi necessária para desfazer equívocos de cálculo pelo fato de o *corpus* conter erros de ortografia, sobretudo nos textos da formação discursiva cotidiana, advindos de redes sociais. Um exemplo: ao buscar no *corpus* pela palavra *araque*, aparecem ocorrências em que se trata obviamente da palavra *ataque* escrita de forma equivocada ("... tentando ver qual seria seu próximo *araque*"). Nesses casos, foi necessário rever os cálculos de IRF para chegar a uma classificação mais correta da palavra.

4 Resultados

Apresentamos a seguir os resultados da análise do CorPS. Dividimos as palavras em dois grupos: as diacríticas e as semidiacríticas. Em cada grupo indicamos as palavras que são de uso geral e as que são "tipicamente" brasileiras ou europeias, isto é, que aparecem em mais de 90% das vezes em textos de uma dessas variedades. Incluímos junto a cada palavra as preposições mais comumente utilizadas.

4.1 Palavras diacríticas

Os quadros seguintes reúnem as palavras que têm um IRF de 100%, ou seja, são consideradas diacríticas.

Quadro 1 – Palavras diacríticas de uso comum nas duas variedades da língua portuguesa.

(de) arromba (de) chofre (a) desoras (às) escâncaras (de) esguelha	(nesse, neste) ínterim (ao) léu (em) prol (de) (de) somenos (à) sorrelfa	(de) soslaio tintim (por) tintim (à) toa (a um, por um) triz (a) trouxe-mouxe

Fonte: elaborado pelos autores

Quadro 2 – Palavras diacríticas de uso típico no português brasileiro.

(de) araque (do) balacobaco (à) beça	(para) dedéu (a) esmo (de) lambuja	(em, aos) pandarecos (à) tona

Fonte: elaborado pelos autores.

Quadro 3 - Palavras diacríticas de uso típico no português europeu.

(às) arrecuas (de) borco	(de) lés (a) lés (na) mouche	(de, em) pantanas

Fonte: elaborado pelos autores.

4.2 Palavras semidiacríticas

As palavras semidiacríticas são as que, em teoria, estão se encaminhando para uma situação de restrição fraseológica completa, mas ainda têm usos não fraseológicos. Essas palavras têm um IRF de pelo menos 80%, isto é, em ao menos 80% de suas ocorrências no *corpus* elas fazem parte de uma estrutura [Prep [SN]] que forma uma locução adverbial ou adjetiva devidamente codificada nos dicionários contemporâneos da língua portuguesa. Os quadros, na sequência, mostram a distribuição dessas palavras entre as duas variedades do português consideradas.

Quadro 4 – Palavras semidiacríticas de uso comum nas duas variedades da língua portuguesa.

(por, ao) acaso	(sob) a égide (de)	(ao) relento
(com) afinco	(sem) eira (nem beira)	(de) repente
(de) antanho	(no, em, ao) encalço (de)	(à) revelia
(com, de) antecedência	(nas) entrelinhas	(em) riste
(de) antemão	(à) espreita (de)	(de cor e) salteado
(em) apuros	(de) estimação	(de) sobejo
(sob os) auspícios (de)	(a, às) expensas (de)	(de) sobreaviso
(às) avessas	(de bom, de mau) grado	(à) socapa
(em) barda	(a) granel	(no, ao) sopé
(na) berlinda	(à) guisa (de)	(em) suma
(aos) borbotões	(por) intermédio (de)	(em, na) surdina
(a, de) bordo	(a) jusante	(a) tiracolo
(de) bruços	(de) manhãzinha	(de) tocaia
(com) brusquidão	(da) matina	(em) torno (de)
(da) carochinha	(à) mercê (de)	(aos) trambolhões
(em, por, para, de) cima	(à) míngua	(para, por, de) trás
(de) cócoras	(de, à) nascença	(nos) trinques
(em, na, à) contraluz	(de, à) noitinha	(de) trivela
(no, em) contrapé	(de) pacotilha	(aos) tropeções
(no) decurso (de)	(à) paisana	(às) turras
(sem mais) delongas	(de) permeio	(em) uníssono
(em) demasia	(em) polvorosa	(por) ventura
(por) desencargo (de consciência)	(no) prelo	(na(s), às) véspera(s)
	(de) rapina	(em) vigor
(por) desfastio	(de) raspão	(em) voga
(em, a) desfavor (de)	(ao, em) redor (de)	
(em) detrimento (de)	(de) relance	

Fonte: elaborado pelos autores.

Quadro 5 – Palavras semidiacríticas de uso típico no português brasileiro.

(de, no) afogadilho	(a) contrapelo	(na) marra
(para o) beleléu	(a) despeito (de)	(de) orelhada
(de, na) butuca	(de uma) figa	(em, de, por) riba
(nos, dos) cafundós	(em) frangalhos	(de) roldão
(fora de) cogitação	(no) frigir (dos ovos)	(de) supetão
(a) contragosto	(na) maciota	(aos) tropeços
(na, em) contramão		

Fonte: elaborado pelos autores.

Quadro 6 – Palavras semidiacríticas de uso típico no português europeu.

(na) alheta	(em, a) catadupa	(sem) desprimor (para)
(à) balda	(às) cavalitas	(aos) magotes
(à, de) borla	(de) chacha	(do) piorio
(da, na) candonga	(ao) desbarato	(de) regadio

Fonte: elaborado pelos autores.

O *script* falhou em identificar palavras que são (semi)diacríticas, mas que são homógrafas de outras palavras de uso livre, como é o caso de *cor* (com vogal aberta na locução adverbial *de cor* = de memória), vão (da locução *em vão* = sem propósito, homógrafa da terceira pessoa do plural do verbo ir) e de *contento* (da locução *a contento* = de modo satisfatório, homógrafa com a forma de primeira pessoa do verbo *contentar*). Também por esse motivo algumas palavras diacríticas foram listadas com um IRF baixo: *arromba*, segundo o dicionário Houaiss (Houaiss; Villar, *online*), é "estilo musical vivo e ruidoso, tocado na viola", e, no *corpus*, tem seu uso totalmente restrito à locução *de arromba* (= sensacional, assombroso, de espantar, estupendo). Apesar disso, ficou com IRF de 80%, no limite do aceitável, por ser homógrafa da terceira pessoa do verbo *arrombar*. Nesse caso conseguimos identificar o equívoco e corrigi-lo, mas pode haver palavras que não foram incluídas na lista por esse motivo.

Um outro problema são as palavras compostas. O *script* identificou muitas delas com alto IRF, sendo algumas diacríticas: (da) *alta-costura*, (em) *alta-voz*, (em) *alto-mar*, (de) *alto-relevo*, (de) *baixo-custo*, (em) *banho-maria*, (de) *bate-pronto*, (ao) *bel-prazer*, (de) *boa-fé*, (de) *bom-tom*, (de) *colarinho-branco*, (a) *curto-prazo*, (do) *deixa-disso*, (ao) *deus-dará*, (em) *larga-escala*, (de) *lesa-pátria*, (a) *longo-prazo*, (de) *mau-gosto*, (a) *meia-haste*, (de) *meia-idade*, (de) *meia-tigela*, (de) *pau-a-pique*, (ao) *pé-coxinho*, (de) *ponta-cabeça*, (a) *queima-roupa*, (à) *tripa-forra*, etc. Não incluímos essas palavras na lista por serem compostas por palavras comuns do português. O composto *trouxe-mouxe* foi incluído por *mouxe* aparecer apenas nesse contexto.

O *script* identificou várias palavras com alto IRF, mas elas não foram incluídas por não estarem codificadas em locuções em nenhum dos dicionários consultados: (de)

aforro, (de) *alterne*, (de) *azinho*, (de) *boaça*, (de) *boinha*, (na) *caruda*, (em) *consonância* (com), (em) *contraciclo*, (com) *denodo*, (ao) *dependuro*, (na) *descontra*, (na) *humilda*, (nas) *imediações*, (pobre de) *marré*, (na) *moralzinha*, (do) *neida*, (à) *rasquinha*, (de) *renome*, (com) *veemência*, (de) *veraneio*. Algumas têm um alto grau de formalidade, como (em) *consonância*, enquanto outras são gírias presentes apenas em textos das formações discursivas cotidiana e literária, como (de) *boaça* (muito de boa) ou (do) *neida* ("do nada", mas dito/ escrito como se a palavra *nada* fosse pronunciada por um falante do inglês).

4.3 Discussão

A aplicação da metodologia proposta ao CorPS permitiu-nos encontrar 137 palavras diacríticas ou semidiacríticas. Como era de se esperar, palavras com restrição fraseológica total são raras, como já apontou o estudo de Holzinger (2018): identificamos apenas 28 palavras diacríticas na análise do *corpus*. As outras 109 são semidiacríticas, com graus variados de restrição fraseológica. Algumas estão quase totalmente restritas, como é o caso de *cima*, com 99,7% de IRF e com usos não fraseológicos muito marginais. Outras ainda são amplamente utilizadas e reconhecidas como palavras autônomas, como *vigor* e *suma* (IRF de 81%), que foram incluídas pelo critério de ter um uso fraseológico maior que 80%. A existência desse *continuum* entre palavras autônomas, semidiacríticas e diacríticas corrobora a hipótese de que haja um processo de restrição fraseológica, com palavras plenamente fossilizadas e outras em transição ao arcaísmo funcionando como índice evidente da evolução do vocabulário (Garcia-Page, 1990).

Em relação à origem das palavras diacríticas portuguesas, a maior parte parece advir de estados arcaicos da língua, isto é, de diacronias passadas do português, por meio da herança latina (*esmo*, *guisa*, *mercê* etc.). Há várias palavras que foram emprestadas de outras línguas, tenha esse empréstimo ocorrido há muito tempo (*soslaio* e *araque*, advindas do espanhol e utilizadas em português pelo menos desde os

séculos XV e XVIII, respectivamente), tenha sido recente (*mouche*, do francês, usada em Portugal a partir do final do século XX). Há ainda palavras diacríticas e semidiacríticas de origem onomatopaica (*chofre* e *triz*, segundo Houaiss; Villar, 2025), palavras originadas de línguas de especialidade (*toa*, termo da náutica) e vários elementos neológicos criados a partir das regras comuns de formação de palavras (*antemão*, *brusquidão*, *carochinha*, *manhãzinha*, *raspão*, *tropeções*, etc.). No que se refere às diferentes palavras diacríticas encontradas tipicamente nas duas variedades do português, destacamos a existência de empréstimos advindos dos substratos indígena e africano no português brasileiro (*butuca*, *cafundós*, *tocaia*). Um fato constante ao buscar pelas etimologias das palavras diacríticas portuguesas nos instrumentos lexicográficos disponíveis é que a maioria é de origem obscura, o que demonstra a necessidade dos estudos históricos.

Por fim, cabe considerar que o único trabalho que apresenta uma lista expressiva de palavras diacríticas do português (Fulgêncio, 2008) reúne apenas 58 itens e não se restringe às ocorrências em locuções adverbiais e adjetivas. Nesse sentido, entendemos que nosso estudo atinge o objetivo de oferecer um levantamento mais amplo e empiricamente fundamentado das palavras fraseologicamente restritas da língua. Entre as palavras diacríticas presentes em locuções adverbiais e adjetivas mencionadas pela autora, nosso *script* deixou de recuperar apenas (às) *pampas*, (às) *mancheias* e (em) *vão*. Embora *pampas* apareça 136 vezes no CorPS, apenas uma ocorrência integra a locução *às pampas*; *mancheias*, por sua vez, surge sempre na locução *a mancheias*, mas tem apenas nove ocorrências no *corpus*, número inferior ao critério de inclusão estabelecido.

5 Considerações finais

Neste artigo buscamos uma caracterização empírica das palavras diacríticas do português a partir de um *corpus* representativo da língua escrita no século XXI. Adotamos uma metodologia de cálculo do Índice de Restrição Fraseológica (IRF)

dessas palavras que demonstrou ser eficaz para identificar termos com alto grau de fixação, o que permitiu observar o *continuum* entre o léxico autônomo e o fossilizado. Os resultados confirmam que as palavras diacríticas formam um grupo reduzido no léxico, mas de grande relevância para a compreensão dos fenômenos fraseológicos em português. A ampla presença de palavras semidiacríticas evidencia que o processo de restrição fraseológica e fossilização lexical é gradual.

Do ponto de vista metodológico, evidenciamos o potencial das ferramentas da Linguística de Corpus para identificar empiricamente os fenômenos tradicionalmente descritos de modo intuitivo. Ao integrar critérios estatísticos e lexicográficos, foi possível propor um modelo de identificação mais rigoroso, passível de replicação e de ampliação a outros recortes do português e a outras línguas românicas.

Em síntese, o estudo pode contribuir para preencher uma lacuna nos estudos fraseológicos do português e oferecer um ponto de partida para investigações futuras voltadas à análise diacrônica dessas palavras, à sua representação lexicográfica e à comparação entre línguas. Acreditamos que, apesar de essas palavras serem relativamente raras e constituírem um fenômeno marginal na língua, são altamente relevantes para compreender os processos de fixação lexical e a transição entre léxico livre e fossilizado (Echenique Elizondo, 2003).

Referências

ACADEMIA DAS CIÊNCIAS DE LISBOA. **Dicionário da Academia de Ciências de Lisboa [online]**. Lisboa: Academia das Ciências de Lisboa, s.d. Disponível em: https://dicionario.acad-ciencias.pt/. Acesso em: 15 set. 2025.

AGUILAR RUIZ, M. J. Vilo, repente y santiamén: los «fósiles fraseológicos» como palabras diacríticas en la fraseología española. *In*: CARMONA YANES, E.; DEL REY QUESADA, S. (org.). **Id est, loquendi peritia**: aportaciones a la lingüística diacrónica de los jóvenes investigadores de la AJIHLE. Sevilla: Universidad de Sevilla, 2011.

AGUILAR RUIZ, M. J. "Fósiles fraseológicos": la configuración formal de "voces fósiles" como palabras idiomáticas en locuciones españolas. **Estudios Humanísticos**. Filología, v. 42, p. 163–183, 2020. DOI https://doi.org/10.18002/ehf.v0i42.6225

ARONOFF, M. **Word Formation in Generative Grammar**. Cambridge: MIT Press, 1976.

AULETE, C. **Dicionário Caldas Aulete [online]**. Rio de Janeiro: Aulete Digital, s.d. Disponível em: https://www.aulete.com.br/. Acesso em: 12 ago. 2025.

BAGNO, M. **Língua, linguagem, linguística**: colocando os pingos nos ii. São Paulo: Parábola Editorial, 2014.

BIDERMAN, M. T. C. Aurélio: sinônimo de dicionário? **ALFA: Revista de Linguística**, São Paulo, v. 44, 2001.

BORBA, F. S. **Dicionário de usos do português do Brasil**. São Paulo: Ática, 2002.

BORBA, F. S. **Organização de Dicionários**: uma introdução à Lexicografia. São Paulo: Ed. UNESP, 2003.

CAMÕES – Instituto da Cooperação e da Língua, I.P. **Dados sobre a língua portuguesa**: Dia Mundial da Língua Portuguesa, 5 de maio de 2023. Lisboa: Camões, I.P., 2023. Disponível em: https://www.instituto-camoes.pt/images/img_agenda2023/Dados_sobre_a_l%C3%ADngua_portuguesa_2023.pdf). Acesso em: 8 set. 2025.

CANIATO, P. C.; SIMÃO, A. K. G. Don Quijote de la Mancha traduzido: unidades fraseológicas diacríticas e suas traduções ao português brasileiro. *In*: ZAVAGLIA, C.; STUPIELLO, E. N. A. (org.). **Tendências contemporâneas dos Estudos da Tradução**. São José Do Rio Preto: Universidade Estadual Paulista, 2015.

ČERMAK, F. *et al*. **Language Periphery**: Monocollocable words in English, Italian, German and Czech. Amsterdam: John Benjamins Publishing Company, 2016. DOI https://doi.org/10.1075/scl.74

ČERMAK, P.; OBSTOVÁ, Z. Palabras monocolocables en español y en italiano: ¿cómo identificar palabras con una colocabilidad muy restringida? **ELUA**: Estudios De Lingüística. Universidad De Alicante, v. 35, p. 53–72, 2021. DOI https://doi.org/10.14198/ELUA2021.35.3

CORPAS PASTOR, G. **Manual de fraseología española**. Madrid: Gredos, 1996.

COSTA, Sérgio Roberto. **Dicionário de gêneros textuais**. São Paulo: Autêntica Editora, 2008.

DOBROVORSKIJ, D.; PIIRAINEN, E. Sprachliche Unikalia im Deutschen: Zum Phänomen phraseologisch gebundener Formative. **Folia Linguistica**, [S. l.], v. 28, n. 3-4, p. 449-473, 1994. DOI https://doi.org/10.1515/flin.1994.28.3-4.449

ECHENIQUE ELIZONDO, M. T. Pautas para el estudio histórico de las unidades fraseológicas. *In*: GIRÓN ALCONCHEL, J. L.; IGLESIAS RECUERO, S.; HERRERO RUIZ DE LOZAIGA, F. J.; NARBONA JIMÉNEZ, A. (org.) **Estudios ofrecidos al profesor José Jesús de Bustos Tovar.** Madrid: Universidad Complutense, 2003.

ECHENIQUE ELIZONDO, M. T. **Principios de fraseología histórica española**. Madrid: Instituto Universitario "Seminario Menéndez Pidal", 2021.

ECHENIQUE ELIZONDO, M. T. Unidades fraseológicas. *In*: DWORKIN, S. N.; CLAVERÍA NADAL, G.; TOLEDO Y HUERTA, A. S. O. **Lingüística histórica del español / The Routledge Handbook of Spanish Historical Linguistics**. Londres: Routledge, 2023. DOI https://doi.org/10.4324/9781003035565-26

EGBERT, J.; BIBER, D.; GRAY, B. **Designing and evaluating language corpora**: a practical framework for corpus representativeness. Cambridge: Cambridge University Press, 2022. DOI https://doi.org/10.1017/9781316584880

FERREIRA, A. B. H. (org.). **Novo Dicionário Aurélio da Língua Portuguesa**: versão eletrônica Século XXI. Rio de Janeiro: Nova Fronteira, 2010.

FULGÊNCIO, L. **Expressões fixas e idiomatismos do português brasileiro**. 2008. Tese (doutorado em Letras) – Pontifícia Universidade Católica de Minas Gerais. Belo Horizonte, 2008.

GARCÍA-PAGE, M. Léxico y sintaxis locucionales: algunas consideraciones sobre palabras "idiomáticas". **Estudios humanísticos. Filología**, León (Espanha), n. 12, p. 279-292, 1990. DOI https://doi.org/10.18002/ehf.v0i12.4052

GARCÍA-PAGE, M. Locuciones adverbiales con palabras "idiomáticas". **RSEL, Revista de la Sociedad Española de Lingüística**, Madrid, n. 21, v. 2, p. 233-264, 1991.

GARCÍA PAGE, M. **Introducción a la fraseología española**. Barcelona: Anthropos, 2008.

GONZÁLEZ REY, M. I. La noción de "hápax" en el sistema fraseológico francés y español. *In*: ALMELA PÉREZ, R.; RAMÓN TRIVES, E.; WOTJAK, G. **Fraseología contrastiva**. Murcia: Univesidad de Murcia, 2005.

HOLZINGER, H. J. Unikale Elemente. Apuntamentos sobre as palabras ligadas fraseoloxicamente do alemán actual. **Cadernos de Fraseoloxía Galega**, Santiago de Compostela, n. 14, p. 165-173, 2012.

HOLZINGER, H. J. Unikale Elemente oder phraseologisch gebundene Wörter? Antworten aus korpuslinguistischer Sicht. **Revista de Filología Alemana**, v. 26, p. 199-213, 2018. DOI https://doi.org/10.5209/RFAL.60149

HOUAISS, A.; VILLAR, M. S. **Dicionário Houaiss da Língua Portuguesa [online]**. Rio de Janeiro: Objetiva / Instituto Antônio Houaiss, s.d. Disponível em: https://houaiss.uol.com.br/. Acesso em: 10 jul. 2025.

LORENZO, J. M. R. **Las palabras diacríticas y sus locuciones en la historia de la lengua española**. 2021. Tese (Doutorado em Estudios Hispánicos Avanzados) - Facultat de Filologia, Traducció i Comunicació, Departamento de Filología Española, Universitat de València, 2021.

PORTO EDITORA. **Dicionário da Língua Portuguesa**. Porto: Porto Editora, 2015.

RUIZ GURILLO, L. **La fraseología del español coloquial**. Barcelona: Editorial Ariel, 1998.

SIMÃO, A. K. G. Lematização de unidades fraseológicas diacríticas em dicionários bilíngues espanhol/português. **Domínios de Lingu@gem**, Uberlândia, v. 8, n. 2, p. 269-288, jul/dez, 2014. DOI https://doi.org/10.14393/DL16-v8n2a2014-15

TEYSSIER, P. **História da língua portuguesa. São Paulo**: Martins Fontes, 2001.

ZULUAGA, A. **Introducción al estudio de las expresiones fijas**. Frankfurt: Peter D. Lang, 1980.

ZULUAGA, A. Phraseologie / Fraseología. *In*: HOLTUS, G.; METZELTIN, M.; SCHIMDT, C. (org.) **Lexikon der Romanistischen Linguistik (LRL).** Band/Volume VI,1. Tübingen: Max Niemeyer, 1992.