# Vidya Syl: A syllable division algorithm for Brazilian Portuguese adhering to the Common Orthographic Vocabulary of the Portuguese Language

**Roges Horacio Grandi[1], Leandro Krug Wives[1], Maria Jose Bocorny Finatto[2], Raquel Salcedo Gomes[1], Bianca Pasqualini[2]**

[1]PPGIE – CINTED – UFRGS
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

roges.grandi@gmail.com, leandro.wives@ufrgs.br, raquel.salcedo@ufrgs.br

[2]PPG-LETRAS-UFRGS – Campus do Vale - Porto Alegre – RS – Brazil

maria.finatto@ufrgs.br, bianca.pasqualini@gmail.com

***Abstract.*** *Syllabification is a fundamental task in natural language processing (NLP) applications, particularly in readability assessment through metrics such as the Flesch Index. This paper presents Vidya Syl, a new syllabification algorithm implemented in Java, developed in compliance with the Brazilian section of the Common Orthographic Vocabulary of the Portuguese Language, known as VOC. The methodological approach consisted of augmenting a rule-based algorithm grounded in phonological and graphical rules with morphological and orthographic rules, until it became compliant with VOC. The experimental results demonstrate improvements in both accuracy and lexical coverage.*

***Resumo.*** *A divisão silábica é uma tarefa fundamental em aplicações de processamento de linguagem natural (PLN), especialmente na avaliação de leiturabilidade por meio de métricas como o Índice Flesch. Este artigo apresenta o Vidya Syl, um novo algoritmo de divisão silábica implementado em Java, desenvolvido em conformidade com a seção brasileira do Vocabulário Ortográfico Comum da Língua Portuguesa, denominada VOC. A abordagem metodológica adotada consistiu em incrementar, com regras morfológicas e ortográficas, um algoritmo baseado em regras fonológicas e gráficas, até que se tornasse aderente ao VOC. Os resultados experimentais demonstram melhorias tanto em acurácia quanto em cobertura lexical.*

## 1. Introduction

Syllable division consists of segmenting words into phonological units called syllables and is essential for applications such as hyphenation, speech synthesis, and reading instruction. In Portuguese, this task is guided by phonological and orthographic principles, such as the mandatory presence of a vowel in each syllable and the specific treatment of vowel and consonant clusters [Câmara Jr. 1992]. Syllable division is a central task in natural language processing (NLP) applications, such as tasks involving estimating the readability of texts using the Flesch index, for example [Flesch 1948].

Despite the existence of rule-based algorithms, many exhibit limitations in covering exceptions and in normative adherence. In this context, the Common Orthographic Vocabulary of the Portuguese Language (VOC) emerges as a standardized lexical resource that enables the validation of syllable divisions according to the official orthography [International Institute of the Portuguese Language 2024].

VOC is the platform that houses the instruments that legally determine the Portuguese orthography. It was officially recognized by the Member States of the Community of Portuguese Language Countries (CPLP) in the final conclusions of the X Conference of Heads of State and Government of the CPLP, held in July 2014 in Dili. In the VOC, it is possible to access the specific dictionary of each member country and to check the frequencies of different forms and syllable divisions used in each country. This work proposes an algorithm that initially relied on phonological rules and progressively incorporated morphological and orthographic constraints, ultimately achieving compliance with the Brazilian section of VOC, namely the *Vocabulário Ortográfico da Língua Portuguesa (VOLP)*, coordinated by Evanildo Bechara [International Institute of the Portuguese Language 2026].

Sections 2 through 7 are organized as follows. Section 2 reviews related work, highlighting rule-based approaches to syllable division in Brazilian Portuguese and discussing their contributions and limitations. Chapter 3 introduces the VOC and its Brazilian section VOLP, which serves as the normative lexical basis for validating syllabification. Chapter 4 presents the Vidya Syl algorithm, detailing its hybrid design, architecture, and innovations aimed at improving adherence to phonological and orthographic rules. Section 5 describes the methodology, including large-scale data extraction from the VOLP, preprocessing steps, baseline evaluation, and incremental rule refinement. Section 6 reports the results, demonstrating the high level of alignment achieved by the proposed method. Finally, Section 7 concludes the study and outlines future research directions.

## 2. Related Works

Silva, Braga, and Resende Jr. (2008) proposed an approach combining two algorithms based on linguistic rules for Brazilian Portuguese: one for syllabification (with 20 rules) and another for stress assignment (with 19 rules), both aimed at applications such as text-to-speech systems. The methods analyze word structure based on vowels and grapheme-to-phoneme context, prioritizing phonological and orthographic principles rather than dictionary-based approaches, which reduces memory usage and allows handling of new words [Silva et al. 2008].

Rocha and Neto (2013) present an algorithm for syllabification in Portuguese that is directly based on the work of Silva, Braga, and Resende Jr. (2008), extending and refining the rule-based approach. The method maintains a focus on phonological and orthographic principles but introduces adjustments to better handle problematic cases, such as consonant clusters, digraphs, and language-specific exceptions. Experimentally evaluated, the algorithm demonstrates good performance and robustness, reinforcing the viability of rule-based approaches for linguistic processing, especially in scenarios that require low computational cost and independence from dictionaries [Rocha and Neto 2013].

Amorim (2024) proposed a rule-based syllabification algorithm for Portuguese implemented in the syllable-pt library, explicitly building upon the works of Silva, Braga, and Resende Jr. (2008) and Rocha and Neto (2013). The approach preserves the linguistic-rule foundation of previous studies while incorporating additional refinements to improve syllable splitting in practical applications [Amorim 2024].

Implemented as a lightweight Java class, the system is designed for easy integration into software projects and operates without reliance on dictionaries, favoring flexibility and low computational cost. As with other rule-based methods, some limitations remain, particularly with irregular or less frequent word patterns; however, the algorithm is designed to evolve, allowing new rules to be incorporated as issues are identified. Distributed under the GNU GPLv3 license, the library represents a practical and extensible continuation of earlier research on Portuguese syllabification.

## 3. The Common Orthographic Vocabulary of the Portuguese Language (VOC) and its Brazilian Section VOLP

The Common Orthographic Vocabulary of the Portuguese Language (VOC) is the platform that hosts the instruments legally defining Portuguese orthography. The VOC has been developed under the coordination of the International Portuguese Language Institute (IILP), the CPLP body responsible for linguistic policy, as mandated by the 2010 Brasília Action Plan for the Promotion, Dissemination, and Projection of the Portuguese Language. Established under the 1990 Orthographic Agreement of the Portuguese Language, signed by CPLP countries, the VOC is being progressively implemented across member states. It gradually integrates the vocabulary of each CPLP country, following political validation and adherence to common methodology and criteria, with representative data from Brazil, Cape Verde, Mozambique, Portugal, and Timor-Leste incorporated to date.

The Brazilian component of the VOC corresponds to the *Vocabulário Ortográfico da Língua Portuguesa*[1], with orthographic information specifically revised and adapted. Adaptations for this specific integration included the proper validation of all listed words and the addition of new information, such as syllabic division, stress marking, inflectional paradigms, functional relationships with other entries (specifically morphologically related words), and frequency indices [Bechara 2017]. This adaptation of VOLP for VOC, was the reference adopted in this study to develop Vidya Syl. This Brazilian section [Bechara 2017], specially developed for the VOC, will henceforth be referred to as VOLP-VOC.

Its database is freely and universally available to citizens, accessible via the internet, containing more than 300,000 entries and over two million orthographic forms. It has been developed in accordance with the latest advances in lexicography and consists of a tagged vocabulary that allows users to identify, among other things, the attested usage of each word in each country, making it possible to determine which words are common to all countries and which occur in specific countries [International Institute of the Portuguese Language 2024].

[1]Orthographic Vocabulary of the Portuguese Language. Our translation.

## 4. Vidya Syl

The Vidya Syl algorithm was developed as an evolution of rule-based approaches for Portuguese syllabification, building primarily on the work of Amorim (2024), which in turn is grounded in the studies of Silva, Braga, and Resende Jr. (2008) and Rocha and Neto (2013). The main goal of Vidya Syl is to advance this line of research by revising and expanding linguistic rules to achieve closer adherence to the VOLP-VOC, ensuring greater consistency with current orthographic standards.

Its implementation was structured in Java 8 with a focus on modularity and efficiency, while preserving independence from dictionary-based resources. One of its key distinguishing features is a right-to-left scanning strategy, which first identifies the vocalic nucleus of each syllable and then determines its surrounding elements, such as adjacent consonants and consonant clusters. This approach enables a segmentation process more closely aligned with the phonological structure of Portuguese, particularly in more complex cases.

The algorithm incorporates refined rules to better handle the patterns codified in VOLP-VOC. As a result, Vidya Syl reduces inconsistencies observed in earlier approaches and improves the accuracy of syllable segmentation in real-world contexts, while preserving the advantages of rule-based methods, such as low computational cost and strong generalization capacity.

### 4.1. Vidya Syl Architecture

Vidya Syl implements syllabification for Brazilian Portuguese using an algorithm based on successive reduction of the word fragment, applying a decision table grounded in phonetic pattern types as proposed by Amorim (2024), with fallback to the first occurrence of certain transitions (vowel–consonant, consonant–vowel, etc.) and handling of special cases [Amorim 2024, Grandi 2026]. It operates on words or texts, returning syllables separated by hyphens. The core logic relies on an extensive set of rules that analyze the phonetic structure of the word. Given a word, after normalization and removal of punctuation marks, hyphens, and special characters, the algorithm converts the word to lowercase. Then, it generates a phonetic pattern based on Amorim (2024), where each letter is mapped to a category:

- v = vowel (a, e, o, á, é, í, ó, ú, ä, ë, ï, ö, ü)
- s = semivowel (i, u)
- t = tilde vowel (ã, ẽ, ĩ, õ, ũ)
- n = nasal consonant (m, n when not followed by h)
- m = mute consonant (b, c, d, f, g, p, t not followed by l, r, t, s, v)
- o = occlusive consonant (b, c, d, p, g, t followed by l, r, t, s, v; also q)
- l = liquid consonant (r, l when not followed by h)
- d = identical double consonant (rr, ss)
- h = digraph with h (lh, nh)
- c = any other consonant

This pattern is used to locate syllable boundaries by applying a series of specific conditions that handle combinations of up to three or four letters at the beginning of the remaining fragment. If no special rule applies, the algorithm searches for the leftmost position of certain sequences (e.g., “cn”, “co”, “sc”, “vc”, etc.).

The core of the algorithm lies in the method `Syllables(String[] words, boolean verbose)`, which iterates over words and, for each one, repeatedly processes a fragment (initially the entire word) until it becomes empty. At each iteration:

- It obtains the pattern of the current fragment.
- It applies a set of rule-based conditions.
- If no condition is satisfied, it attempts to find the lowest index among several substrings (e.g., "cn", "co", "sc", "vc", "lsc", "nsl", etc.) and splits the fragment accordingly; otherwise, it adds the entire fragment as a syllable.

The algorithm includes hundreds of specific rules, organized into methods named after combinations of phonetic patterns. Each method analyzes the pattern to determine how many letters should be separated as a syllable. The separation is performed by removing the initial characters of the fragment and adding them to the syllable list.

A set of special cases is explicitly addressed, including:

- Non-divisible words: a predefined set of words, initialisms, and acronyms, such as "IPVA", "jeans", and *"yin"*, that are not split.
- Roman numerals: they are preserved as whole units.
- Vowel constraints: if the fragment does not contain at least two vowels, the word is considered indivisible (with exceptions such as "brunch" and "mach").

The algorithm is highly heuristic, covering exceptions and loanwords (e.g., shock, week, brunch), and prioritizes specific rules over the general principle of splitting before a consonant that begins a non-permitted cluster. Its complexity is justified by the need to ensure adherence to the syllabification rules established by the VOLP-VOC.

## 5. Methodology

This section describes the methodological procedures adopted to develop and evaluate the Vidya Syl algorithm. It includes the large-scale extraction of syllabified entries from the VOLP-VOC, followed by data cleaning steps such as the removal of polylexical lemmas and duplicate records. Subsequently, an initial evaluation was conducted using the baseline algorithm proposed by Amorim (2024), and an iterative process of rule refinement was carried out to maximize alignment with the VOLP-VOC. The section concludes with the criteria used to determine the final level of adherence achieved by the proposed method.

### 5.1. Retrieval of syllables from the VOLP-VOC

The Vidya Syl class `MineVolpEntries.java` was employed to retrieve all syllabifications available in the VOLP-VOC, which are accessed through lemma queries on the web via Uniform Resource Identifiers (URIs), as illustrated in Figure 1. Given that the highest lemma reference in the VOLP-VOC is 241,289, it systematically queried the entire range of lemma identifiers, from 1 to 241,289.

In this process, 224.238 entries were successfully returned; that is, the search and identification algorithm for the syllables of 17.051 lemmas generated some computational exceptions that could not be processed. For instance, when `n = 1`,

VOLP-VOC returns the content displayed in Figure 2: the verb *cultivar* segmented into the syllables by points.


**Figure 1. URI for lemma queries in the VOLP-VOC.**

The next step consisted of converting the VOLP-VOC notation of syllabification and lexical stress into the hyphenated form adopted by Vidya Syl. This transformation was performed by the class `AdjustVolpLemmas.java`. For instance, the notation `(cul.ti.'var)` was converted to `cul-ti-var`.


**Figure 2. Return of lemma 1 from the VOLP-VOC.**

### 5.2. Removal of Multiword, Unattested, and Duplicate Lemmas

As discussed by Nunes (2022), polylexicality occurs when a lexical unit (defined as a word or a sequence of words with a single meaning that is treated as an element of a language's lexicon) is composed of at least two lexical elements (words). The query to the VOLP-VOC lemmas returned 43 polylexical lemmas, which were removed from the lemmas set [Nunes 2022].

For 19 words, the VOLP-VOC provides different syllabifications depending on the grammatical class. These words are: *adônio, afidídeo, aguaiano, ajuizador, antidiurético, atribuidor, banyamulenge, braille, carajá-mirim, chauffeur, haeckeliano, iao, insetófilo, junkie, macauense, picuipeba, saudar, wronskiano,* and *zombie*. In such cases, one syllabification per word was selected at random. From this process, it was removed 24 entries.

Following manual revision via the VOLP-VOC query interface, five additional entries flagged as not attested were removed. These words are: *fröbeliano, ci-iao, fecho-éclair, hors-d´oeu-vre,* and *codo-d´áua*.

Many words were registered under more than one grammatical class while sharing the same syllabification. The word *tal*, for instance, can function in several distinct grammatical classes, as shown in Figure 3. After disregarding duplicate entries corresponding to words that share the same syllabification across distinct grammatical classes, 186,241 unique words and their respective syllabifications remained.

### 5.3. Initial Evaluation at Amorim's Code

Before initiating the process of incorporating phonological and orthographic rules to align Vidya Syl with the VOLP-VOC, an evaluation was conducted on the selected


**Figure 3. Different classes of word *tal* in the VOLP-VOC.**

words and their corresponding syllabifications using Amorim's original algorithm. In this initial phase, it was found that 49,467 (26.56%) of the 186,241 available syllabic divisions were divergent.

### 5.4. Incremental Insertion of Morphological and Ortographic Constraints in Vidya Syl

An extensive process of incorporating phonological and orthographic rules into Vidya Syl was then initiated. After each new rule was added, the alignment rate with the VOLP-VOC was recomputed to determine whether it had increased or decreased. All 186,241 entries were verified by the Vidya Syl class `CheckSyllabesVolp.java`, with each iteration taking approximately 13 seconds. When alignment improved, the rule was retained; when it decreased, the rule was discarded.

### 5.5. Finalization

Figure 4 shows the final iteration of improvements, in which 186,186 entries were marked as `ok` (i.e., compliant with the VOLP-VOC) and 55 entries as `wrong` (i.e., non-compliant with the VOLP-VOC).

```
partial: 00:00:01       total:10000     ok: 10000       wrong:0
partial: 00:00:02       total:20000     ok: 19998       wrong:2
partial: 00:00:02       total:30000     ok: 29998       wrong:2
partial: 00:00:03       total:40000     ok: 39997       wrong:3
partial: 00:00:04       total:50000     ok: 49995       wrong:5
partial: 00:00:05       total:60000     ok: 59995       wrong:5
partial: 00:00:05       total:70000     ok: 69993       wrong:7
partial: 00:00:06       total:80000     ok: 79993       wrong:7
partial: 00:00:07       total:90000     ok: 89990       wrong:10
partial: 00:00:07       total:100000    ok: 99987       wrong:13
partial: 00:00:08       total:110000    ok: 109986      wrong:14
partial: 00:00:09       total:120000    ok: 119983      wrong:17
partial: 00:00:10       total:130000    ok: 129982      wrong:18
partial: 00:00:10       total:140000    ok: 139980      wrong:20
partial: 00:00:11       total:150000    ok: 149977      wrong:23
partial: 00:00:11       total:160000    ok: 159965      wrong:35
partial: 00:00:12       total:170000    ok: 169954      wrong:46
partial: 00:00:13       total:180000    ok: 179952      wrong:48
VOLP
final.: 00:00:13        total:186241    ok: 186186      wrong:55
```

**Figure 4. URI for lemma queries in the VOLP-VOC.**

Considering the 186,241 syllabifications retrieved from the VOLP-VOC, the alignment process was concluded when Vidya Syl reached 99.97% agreement with the reference (186,191 out of 186,241 entries). The remaining 55 words could not be aligned due to uncertainties regarding their correct syllabic representation in the VOLP-VOC itself. The verification of these 55 uncertain cases was deferred to future work, to be carried out jointly with linguists specialized in phonology and lexicography:

1. Table 1 lists six words for which doubts remain regarding the separation of double consonants;

2. Table 2 lists eleven words for which uncertainties persist regarding the syllabic segmentation of the falling diphthong *au*;
3. Table 3 lists four words for which uncertainties persist regarding the syllabic segmentation of hiatuses;
4. Table 4 lists nine words with doubts regarding Portuguese orthographic rules;
5. Table 5 lists three words for which uncertainties persist regarding the syllabification of English loanwords; and
6. Table 6 lists 22 words whose syllabic division conflicts with similar ones in the VOLP-VOC.

**Table 1. Words with doubts regarding the separation of double consonants.**

| Word | VOLP-VOC Syllables | Word | VOLP-VOC Syllables |
|---|---|---|---|
| antisséptico | an-tissép-ti-co | rickettsiácea | ric-kett-si-á-cea |
| rickettsíase | ric-kett-sí-a-se | rickettsiáceo | ric-kett-si-á-ceo |
| rickettsiose | ric-kett-si-o-se | wölffliniano | wölffli-ni-a-no |

**Table 2. Words with doubts regarding the segmentation of the falling diphthong *au*.**

| Word | VOLP-VOC Syllables | Word | VOLP-VOC Syllables |
|---|---|---|---|
| ressaudação | res-sa-u-da-ção | saudador | sa-u-da-dor |
| ressaudado | res-sa-u-da-do | saudável | sa-u-dá-vel |
| ressaudar | res-sa-u-dar | saudosismo | sa-u-do-sis-mo |
| saudação | sa-u-da-ção | saudosista | sa-u-do-sis-ta |
| saudade | sa-u-da-de | saudoso | sa-u-do-so |
| saudado | sa-u-da-do | | |

**Table 3. Words with doubts regarding the separation of double consonants.**

| Word | VOLP-VOC Syllables | Word | VOLP-VOC Syllables |
|---|---|---|---|
| coesposo | coes-po-so | reengineering | reen-gi-nee-ring |
| sacroilíaco | sa-croi-lí-a-co | yeti | ye-ti |

**Table 4. Words with doubts regarding Portuguese orthographic rules.**

| Word | VOLP-VOC Syllables | Word | VOLP-VOC Syllables |
|---|---|---|---|
| araquãí | a-ra-qu-ã-í | reganho | re-gan-ho |
| icozeiro | i-cozei-ro | tubinar | tu-bi-na-r |
| malcoado | malco-a-do | tupperware | tup-pe-rwa-re |
| quão | qu-ão | westphalense | westp-ha-len-se |
| reaquisto | re-a-qui-sto | | |

**Table 6. Words involving uncertainty regarding similar terms in the VOLP-VOC.**

| Word | VOLP-VOC Syllables | VOLP-VOC Similar |
|---|---|---|
| business | busi-ness | `("co.re-bu.'si.ness)` |
| contraemergentes | con-trae-mer-gen-tes | `(con.tra.e.mer.*gen.te)` |
| diretora-geral | di-re-tora-ge-ral | `(di.re.'to.ra)` |
| email | e-ma-il | `("e-'mail)` |
| físico-quimicamente | fí-si-co-qu-imi-ca-men-te | `(qui.mi.ca.'men.te)` |
| funboard | fun-bo-ard | `(clip;'board)` |
| geográfico-linguisticamente | ge-o-grá-fi-co-lin-gu-isti-ca-men-te | `(lin.guis.ti.ca.'men.te)` |
| heautognóstico | he-au-tog-nós-ti-co | `(heau;to.'gno.se)` |
| keynesiano | ke-y-ne-si-a-no | `(key;ne.si.a.'nis.mo)` |
| magnificat | ma-gni-fi-cat | `(mag.ni.fi.'ca.do)` |
| miudinha | miu-di-nha | `(mi.u.*di.nho)` |
| neomalthusianista | ne-o-malt-hu-si-a-nis-ta | `(mal;thu.si.'a.no)` |
| pipuiruçu | pi-pui-ru-çu | `(pe.pu.i.ru.*çu)` |

| Word | VOLP-VOC Syllables | VOLP-VOC Similar |
|---|---|---|
| pituitária | pi-tui-tá-ria | `(pi.tu.i.*tá.ri,o)` |
| poliuretana | po-liu-re-ta-na | `(po.li.u.re.*ta.no)` |
| rato-da-índia | ra-to-da-ín-di-a | (’ín.di,a) |
| sã-luisense | sã-lu-i-sen-se | `(lui.*sen.se)` |
| são-luisense | são-lu-i-sen-se | `(lui.*sen.se)` |
| serial-killer | se-rial-kil-ler | `(se.ri.’al)` |
| suicidomaníaco | sui-ci-do-ma-ní-a-co | `(su.i.*ci.da)` |
| tapuirana | ta-pui-ra-na | `(a.pu.i.*ra.na)` |
| wesleyanismo | wes-ley-a-nis-mo | `(wes.le.y.a.*nis.ta)` |

## 6. Results

The 99.97% alignment achieved with the VOLP-VOC was considered a final point for the first version of Vidya Syl. When consulting linguistics and phonology specialists from the Linguistics Department of UFRGS, they were positive about the work, but had difficulty confirming the accuracy of the syllabic divisions provided by the VOLP-VOC for the 55 uncertain words. This validation may be revisited at a later date.

### 6.1. Impact on Flesch Reading Ease (FRE) Index

The Flesch Reading Ease (FRE) is a widely used readability metric that evaluates textual comprehensibility based on average sentence length and average word length (in syllables). It yields scores from 0 to 100, where higher values indicate easier texts and lower values denote greater complexity. Originally proposed by Rudolf Flesch, it is broadly applied in education and natural language processing [Flesch 1948].

The study conducted by Martins et al. (1996) focuses on the adaptation of the FRE Index to Brazilian Portuguese, translating and adjusting the original formula to reflect the specific linguistic characteristics of the language, and proposing the following adapted equation [Martins et al. 1996]:

$$\textbf{Index} = 248.835 - 1.015 \times \left(\frac{\text{words}}{\text{sentences}}\right) - 84.6 \times \left(\frac{\text{syllables}}{\text{words}}\right) \tag{1}$$

The Portuguese text analyzed by the tools in this evaluation:

> *Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a ideia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta.*

Using the equation proposed by Martins et al. (1996), along with a dedicated syllabification algorithm, NILC Metrix NLP Tool [Venturus 2024] found 3 sentences, 69 words, 23 words by sentence and 2,31 syllables by word. Using these parameters, the resulting FRE Index score was **29.3**.

Using the same equation proposed by Martins et al. (1996), Vidya Text [Grandi et al. 2026], an Educational Tool that employs Vidya Syl as its syllabification algorithm, found 3 sentences, 68 words, 22,7 words by sentence and 2,3 syllables by word. Using these parameters, the resulting FRE Index score was **31.7.**

The different results obtained with the same equation, using different syllabification algorithms, highlight the impact of this factor on the computation of readability indices.

**Table 5. Words with doubts regarding the syllabification of English loanwords.**

| Word | VOLP-VOC Syllables | Word | VOLP-VOC Syllables |
|---|---|---|---|
| washingtónia | was-hing-tó-nia | washingtônia | was-hing-tô-nia |
| washingtonite | was-hing-to-ni-te | | |

## 7. Conclusions and Future Works

This study presented Vidya Syl, a syllable division algorithm for Brazilian Portuguese, designed to achieve high adherence to the VOLP-VOC. Building upon prior rule-based approaches, particularly the algorithm proposed by Amorim, the proposed method introduced a hybrid strategy that combines phonological and orthographic rules with large-scale lexical validation. This combination proved effective in addressing known limitations of purely rule-based systems, especially with regard to exception handling and normative consistency.

The results demonstrate that Vidya Syl achieves a high level of alignment with the VOLP-VOC, reaching 99.97% agreement over a dataset of 186,241 lexical entries. This represents a significant improvement over the baseline algorithm, which exhibited 26.56% divergence in the initial evaluation. The incremental rule insertion strategy, guided by empirical validation against the VOLP-VOC, was instrumental in achieving this level of performance, while preserving the advantages of rule-based methods, such as low computational cost, interpretability, and independence from large lexical databases at runtime.

Despite these advances, the analysis also revealed inherent challenges in the standardization of syllabification. A small subset of 55 words could not be conclusively aligned due to uncertainties in their syllabic representation within the VOLP-VOC itself, even when evaluated by domain experts. This finding highlights not only the complexity of syllabification in Portuguese but also potential limitations or ambiguities in normative lexical resources.

Future work will focus on many directions: i) to rely on expert judgment to settle the 55 words whose syllabification remained inconclusive; ii) efforts will be made to optimize the algorithm; and iii) integrate it into NLP applications, such as readability tools and educational technologies. Its code will be made available for these purposes and also to enable other research groups to carry out similar work with other Portuguese dialects in relation to the VOLP-VOC.

In summary, Vidya Syl represents a robust and extensible contribution to Portuguese NLP, demonstrating that hybrid approaches grounded in both linguistic theory and lexical validation can achieve high levels of accuracy and practical relevance. Furthermore, its results already feed into a first version of the Vidya Text Tool, which indicates the readability potential of a text according to the Brazilian Flesch index formula, but now with a new index in the syllable count.

## 8. Acknowledgments

This research was supported by CNPq (National Council for Scientific and Technological Development), grant PQ 307088/2023-5, National Institutes of Science and Technology – CNPq-TILD-IAR (grant 408490/2024-1), PROBIC FAPERGS-

UFRGS, and FAPERGS Ed. 06/2025 - PROGRAMA PESQUISADOR GAÚCHO – PqG (grant 25/2551-0002702)

## 9. Ethics Statement

No sensitive personal data were used in the development of the Vidya Syl algorithm. The research is aligned with the ethical principles of the field of Computational Linguistics and aims to provide the community with a properly tested syllabification algorithm.

## 10. Limitations

A limitation encountered during this study is that VOLP-VOC occasionally assigns different syllabifications to the same lexical item according to its grammatical category. For example, braille is represented as a masculine noun ('brail;le) and as an adjective (bra.il.le). A further limitation concerns the verification of correct syllabification across the full inventory of Brazilian Portuguese lexical entries. Despite this challenge, the proposed algorithm achieved full agreement with the syllabifications provided in the VOLP-VOC, except for 55 words whose correct syllabification remains uncertain.

## 11. Statement on the Use of Generative AI

In accordance with CNPq Ordinance No. 2,664/2026, the use of the GPT-5.5 model of ChatGPT is hereby declared as a support tool for stylistic refinement and for translating the original text from Portuguese into English. The authors are fully responsible for the final content presented.

## References

Amorim, E. (2024). syllable-pt: Portuguese syllabification algorithm [source code]. `https://github.com/evelinamorim/syllable-pt`.

Bechara, E., editor (2017). *Vocabulário Ortográfico da Língua Portuguesa*. Academia Brasileira de Letras, Rio de Janeiro, 5 edition. Edição adaptada ao VOC.

Câmara Jr., J. M. (1992). *Estrutura da língua portuguesa*. Vozes.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Grandi, R. H. (2026). Vidya syl: A syllable division algorithm adhering to the brazilian section of the common orthographic vocabulary of the portuguese language (volp). `https://gitlab.com/rogesgrandi/vidyasyl`. Accessed: 19 Jul. 2026.

Grandi, R. H. et al. (2026). Vidya text: an educational tool to support the teaching of writing and the study of brazilian portuguese vocabulary. https://vidyatext.nuvem.ufrgs.br. Acesso em: 19 jul. 2026.

International Institute of the Portuguese Language (2024). Voc: Vocabulário ortográfico comum da língua portuguesa. `https://voc.cplp.org/index.php?action=von&von=br`.

International Institute of the Portuguese Language (2026). Vocabulário ortográfico comum da língua portuguesa. https://voc.cplp.org/index.php?action=ficha. Comunidade dos Países de Língua Portuguesa (CPLP).

Martins, T. B., Ghiraldelo, C. M., Nunes, M. d. G. V., and Oliveira Junior, O. N. d. (1996). Readability formulas applied to textbooks in brazilian portuguese.

Nunes, Q. O. B. (2022). *Unidades fraseológicas em Sagarana de Guimarães Rosa: Um estudo de traços culturais em tradução ao francês*. Tese de doutorado, Universidade Federal de Mato Grosso do Sul.

Rocha, W. and Neto, N. (2013). Implementação de um separador silábico gratuito baseado em regras linguísticas para o português brasileiro. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*.

Silva, D. C., Braga, D., and Jr., F. G. V. R. (2008). Separação das sílabas e determinação da tonicidade no português brasileiro. In *XXVI Simpósio Brasileiro de Telecomunicações (SBrT)*, pages 1–5.

Venturus (2024). Nilc-metrix. https://pln.venturus.dev.br/nilcmetrix. Accessed: 6 abr. 2026.