Improving the portuguese hyphenation rules

Leonardo Araújo Aline de Lima Benevides

June 14, 2024

Abstract

Portuguese hyphenation rules are available for more than 35 years and have done a good job. Nonetheless they still make mistakes and leave some hyphenation points unmarked. Although most undetected hyphenation points are located near word boundaries, what will be irrelevant for TEX typographic purposes, they are still useful to hyphenate proper nouns, new words or pseudoword, and for usage in other applications, such as text-to-speech conversion. A list of 85 638 hyphenated words acquired from online dictionaries was used along *patgen* to create improved rules, leading to better hyphenation of Portuguese words.

1 Introduction

Hyphenation in text wrapping was not used for a long time. Words should fit entirely in a line, or they would be broken in arbitrary places. Initially, no markers were used to indicate word wrapping, leading to potential confusion and unintended interpretations. As a result, orthographers advocated for the introduction of a sign to indicate such breaks. Portuguese faced the same gradual introduction of a hyphenation sign to mark words wrapping across lines. Even though the usage of a hyphenation sign (=) was advocated by orthographers (Magalhães Gândavo 1574), few documents used such sign until the end of the 18th century (Araújo and Maruyama 2015).

In some cases, hyphenation hinders smooth reading and should be avoided in child literature. In opposition, both excessive spacing and insufficient spacing between words also impose difficulty in the reading process, making hyphenation fundamental when texts use short line lengths. As a line gets shorter, the number of breaking candidates between words decreases, leading to awkward spaces between words and among letters. For that matter, automatic hyphenation plays an important role in good typesetting.

TEX is a typesetting system which carefully deals with these issues, automatically arranging text on a page to create a good reading experience. Automatic hyphenation is an important part on this process, promoting an even-tempered distribution of elements on the page. Line height and line length, paragraph length, font size and typeface, letter-spacing, and word-spacing are some factors which influence the text legibility and readability. Space between words should not be too long creating lakes and rivers in the text, nor too tight, impairing the legibility and readability.

Another important matter to consider is ambiguities that might be created when a word is partitioned during the hyphenation process. In English we should avoid hyphenations such as *re-cover*, *re-form*, *re-sign*, *the-rapist*, *depart-mental*, and *mans-laughter* (in Portuguese, some examples are: *de-putada*, *fede-ração*, *acumula*, *após-tolo*, *cú-bico*). Hyphenations that might lead the reader to pronounce a word incorrectly should also be prevented. That is the case of *considera-tion*, in Enslish (and *pe-rigo* in Portuguese).

In some situations, hyphenation is also a matter of style. Some partitioning choices sound better than others. These conflicting alternatives typically arises when a words has many possible hyphenation points. Consider *ar-chae-ol-o-gist* (*or ar-che-ol-o-gist*), which is preferably partitioned as *archae-ologist* (or *arche-ologist*) in opposition to *archaeol-ogist* (or *archeol-ogist*) or *archaeolo-gist* (or *archeolo-gist*). It is preferable to keep whole morphemes together. In the previous example: *archae* (or *arche*, meaning "ancient", "primitive") and *-ologist* ("one who studies the topic"). In Portuguese, it is also more elegant to avoid splits between double consonants or vowels, even if an hyphenation point do exists between those letters. For example, *pressu-rizar* is preferable over *pres-surizar* and *empreen-dedor* is preferable over *empre-endedor*. Even so, exceptions exists, it is preferable to partition *micro-organismo* rather than *microor-ganismo* (keeping morphemes together is favored over splitting a double vowel). Numerous factors come into play when choosing the optimal hyphenation point for a word.

The general rule for Portuguese hyphenation is to split a word into its syllables. A syllable is made of a mandatory nucleus, filled by a vowel, and optional peripheral consonants (before or after the nucleus). In some situations, the syllabic division does not respect the ethnologic constituents. The usage of the prefixes *bis-* and *in-* are examples of this circumstance. The correct syllabifications are *bi-sa-vô*, *i-nobs-tan-te*[1], and *i-na-ti-vo*, where the prefixes are split into two syllables. But, as pointed out previously, it is preferable to

[1]The rule of syllable division could lead to two possible partitions: *i-nobs-tan-te* and *in-obs-tan-te*, but the first is preferable.

keep morphemes together rather than splitting them apart. Therefore we should favor the hyphenation *bisa-vô* over *bi-savô*, *inobs-tante* over *i-nobstante*, and *ina-tivo* over *i-nativo*. This last example could also lead to a misunderstanding since the word *nativo* (*native* in English) emerges from this word break.

Each language has its hyphenation rules, which can be categorized into two groups: those driven by morphology (etymology) and those driven by pronunciation. An algorithmic approach employs a logical system to analyze words and apply the hyphenation rules of a specific language. Since hyphenation rules vary significantly between languages, an algorithm must be developed for each one. While a logic-based system can be efficient and compact, it still needs to address exceptions through hard-coded rules.

The automation of this process can employ various approaches, including:

Dictionary-Based Approach: This method restricts hyphenation possibilities to entries found in the dictionary.

Algorithmic-Based Approach: This approach can be applied to any sequence encountered in a text.

Rule-Based Model: This model recognizes prefixes, suffixes, morphemes, or specific sequences suitable for hyphenation.

Pattern Matching: By using a corpus of hyphenation examples in a language, this approach identifies letter sequences that determine suitable hyphenation points. Patterns encompass prefixes, suffixes, exceptions, and special hyphenation rules of the language.

Mixed Approach: This approach combines two or more of the previously described methods to enhance hyphenation accuracy and flexibility.

Analyzing only the immediate surroundings may not always suffice to determine a potential hyphenation point. For instance, consider *de-moc-ra-cy* and *dem-o-crat*, where the immediate surrounding of the letter *e* does not provide a clear indication of the hyphenation point. Even when facing chosen patterns that typically make hyphenation straightforward, exceptions can arise. Take the sequence *tion*, for example. It may seem logical to place a break before this pattern, but the word *cation* is hyphenated as *cat-ion*, highlighting the influence of etymology on the way a word is split.

A word with multiple meanings may have distinct hyphenations based on its intended meaning. 'For instance, the Swedish word form *glassko* has three different meanings, and can be hyphenated as *glas-sko* (glass shoe), *glass-ko* (ice cream cow) and in the non-standard way, *glass-sko* (ice cream shoe)' (Németh 2006). In Portuguese, the word *sublinha* might be hiphenated in two different ways: *su-bli-nha* when representing the inflected form of the verb *sublinhar* (to underline) or *sub-li-nha* when refering to the line under (underline as a noun).

Machine learning techniques were employed to hyphenate Norwegian text (Kristensen and Langmyhr 2001). The study revealed that, overall, the TEX approach outperformed a neural network. Both methods demonstrated similar performance in identifying correct hyphenation points and minimizing incorrect hyphenations when tested on a small word list (with the neural network slightly outperforming in correctly recognizing hyphenation points). However, when tested on a larger word list, the TEX approach proved superior in avoiding incorrect hyphenations compared to the neural network.

The original TEX hyphenation algorithm, introduced by Knuth (1977), primarily focused on the English language and employed three main rules: (1) suffix removal; (2) prefix removal; and (3) the Vowel-Consonant--Consonant-Vowel (VCCV) breaking rule[2]. Tests demonstrated that it could identify 40% of allowable hyphen locations (F. M. Liang 1983). The TEX hyphenation algorithm later adopted the approach proposed by Frank M. Liang, which involves competing patterns. The TEX82 algorithm employs five alternating levels of hyphenating and inhibiting patterns. The program for pattern generation, known as PATGEN, was created by (F. Liang and Breitenlohner 1991) and has been utilized to generate hyphenating patterns for numerous languages (P. Sojka 1995b; P. Sojka 1995a; P. Sojka et al. 2005; P. Sojka and Antoš 2003; Scannell 2003). It involves sweeping a database of hyphenated words in a language to identify both hyphenating and inhibiting patterns, ultimately creating a list of competing patterns for that specific language.

The effective hyphenation of words by TEX will actually depends on the following factors: (1) document language, which will determine which set of patterns to apply; (2) characters used, since some might block hyphenation at their edges; (3) the value of the internal variables `lefthyphenmin` and `righthyphenmin`[3], which defines the minimum sequence length of characters at the left and right borders before any hyphenation is allowed.

[2]The VCCV rule in hyphenation patterns, places the syllable boundary between two consecutive consonants when they appear between two vowels. For example, in the word *sudden*, the syllable break occurs between the *d* and the second *d*, making it *sud-den*. The VCCV pattern is a relatively common syllable division pattern in English. This rule ensures accurate hyphenation, maintaining word readability and pronunciation when words are split at the end of a line in printed text.

[3]The variables `lefthyphenmin` and `righthyphenmin` are language dependant and are defined in *tlpobj* files (`/usr/local/texlive/20XX/tlpkg/tlpobj/hyphen-xxxxxx.tlpobj`). Default values varies in the range from 1 to 3. English and Portuguese, for example, use `lefthyphenmin=2` and `righthyphenmin=3`.

Despite the fact that TEX hyphenation algorithm and rules are old, they are, to these days, the most frequently used approach, even outside the TEX's world. The grounds for this is Hunspell, a spell checker and morphological analyzer that is adopted in many softwares (e.g. LibreOffice, OpenOffice.org, Mozilla Firefox, Mozilla Thunderbird, Google Chrome, macOS, InDesign, memoQ, Opera, Affinity Publisher, among others (Hunspell's Team 2023a)). Hunspell uses TEX hyphenation rules (Hunspell's Team 2023b; Levien 1998), making TEX hyphenation widespread in the computer world. That is a result of TEX approach simplicity and versatility. The algorithm works effectively, as it already supports rules for 66 languages (TEX pattern authors 2023), and offers the flexibility to create rules for any currently unsupported languages.

Unfortunately, certain hyphenation rules cannot be implemented using the TEX hyphenation algorithm. For example, in German, hyphenation can lead to letter change or insertion. Additionally, compound words lack hyphens, resulting in extended letter sequences without visible separation and even repetitions of the same letter, as seen in examples like *Wasserrinne* and *Schifffahrt*. Furthermore, the German spelling reform made some changes, making it necessary to create a different set of rules for German hyphenation. For example, the word *Schiffahrt* should be hyphenated as *Schiff-fahrt*, preserving the *f*s from each word that makes this compound. The hyphenation should insert an *f* that is not part of the written form. That was not a problem for the old written form of the word: *Schifffahrt*. Also, the old hyphenating rules of German grammar stated the hyphenation *Bäk-ker* for the word *Bäcker*, *Zuk-ker* for the word *Zucker* and *pak-ken* for the word *packen*. Nowadays, those words are hyphenated as *Bä-cker*, *Zu-cker* and *pa-cken*, respectively. Some words have also changed their hyphenation point after the spelling reform. For example, *Fen-ster* became *Fens-ter* and *mei-stens* became *meis-tens*. Some problems in compound word hyphenation in TEX are discussed in P. Sojka (1995a).

Moreover, as mentioned earlier, certain hyphenations may be preferred for stylistic reasons, or to avoid ambiguity, or for better reading experience. Some words might have multiple hyphenations, depending on the intended meaning. In such cases, TEX may not efficiently address the hyphenation challenge, as it would necessitate case-by-case handling.

2 Patterns for TEX hyphenation

To simplify, if we consider only the Latin alphabet, with no diacritcs, the patterns used in TEX hyphenation are of the form: `^\.?[0-9]?([a-z]+[0-9]?)+\.?$`, where we have described it using a regular expression[4]. One example of such pattern is `4z1z2`, which is composed of a sequence of letters and numbers. In general, we use characters/symbols from the language alphabet, along with Hindu-Arabic numbers, to express either hyphenation facilitation or inhibition. Odd numbers indicate a good hyphenation point, whereas even numbers indicate a bad place to break. The given example states that the sequence has a good breaking point between the first and the second *z* and an hyphenation should be inhibited before the first *z* and after the second *z*. For example, the hyphenation of the word *piz-za*, *fiz-zle* and *mez-zanine* use this rule, where we see the hyphen placed between the two *z*'s and no hyphen before, nor after the *z*'s. Patterns may also use period symbol (.) to indicate word boundaries. The pattern `.sh2` applies to beginning of words, implying that the *s* and the *h* should stick together in beginning of a word and an hyphenation should also be inhibited after the *h*. For example, this pattern is used in *Sher-lock*.

Hyphenation rules are organized in levels, from 1 to 9, where odd numbers represent hyphenating levels and even numbers represent inhibiting levels. Each level works as an exception level of it predecessor. For example, the rule `sh1er` indicate a good hyphenation point between the *h* and the *e* in the sequence *sher*. A rule at a higher level, as `.sh2`, implies an exception to the lower level rule. When we see *sher*, in the beginning of a word, the rule `.sh2` applies and hyphenation proposed by the lower level rule `sh1er` should be hindered. That is the case in the hyphenation of the word *Sher-lock*. The full example is provided in Listing 1, where we might see all pertinent English rules taking place in the hyphenation of *Sher-lock*.

[4]Regular expressions (regex) are powerful search patterns used in text processing to find, match, and manipulate strings of text. They are a fundamental tool in programming and are supported in many programming languages. The regex given here uses the Perl syntax and might be broken in the following parts: `^` and `$` mark the start and end of the string; `\.?` specifies an optional period; `[0-9]?` allows for an optional single digit; `([a-z]+[0-9]?)+` matches sequences of lowercase letters interleaved with optional digits.

Listing 1: Example of rules applyied in the hyphenation of the word *Sherlock*. Example done using a port of TEX's hyphenation algorithm to Go provided at `https://github.com/speedata/hyphenation`.

```
    .   s   h   e   r   l   o   c   k   .
        0   0   2   |   |   |   |   |   |     .sh2
        0   2   0   |   |   |   |   |   |     s2h
        0   0   1   0   0   |   |   |   |     sh1er
        |   |   |   0   1   0   |   |   |     r1l
        |   |   |   0   3   0   4   |   |     r3lo4
        |   |   |   |   |   |   0   0   1     ck1
max:    0   2   2   0   3   0   4   0   1
final:  s   h   e   r   -   l   o   c   k   -
```

In summary, a pattern will consist of a string made of characters (from the language alphabet) possibly with a number in between, expressing the hyphenation/inhibition level. Occasionally word boundaries marker (the period) is used at the pattern edges. When there is no number between characters in a pattern, a zero is assumed, which means *undefined* and no hyphenation point will be suggest at that location.

3 Patgen

Patgen utilizes a list of hyphenated words to extract patterns and use them to define rules at various levels and lengths. It starts with short patterns and incrementally increases their length until reaching the maximum pattern length allowed by the user. The objective is to keep the patterns as concise as possible, as this enhances their generalizability. As it advances and incorporates longer patterns, *patgen* establishes exceptions. In certain cases, analyzing long patterns may be necessary, as some hyphenation points could depend on characters far away from the breaking point[5].

Patgen works on glyph[6] indices rather than character codes. Each glyph is represented by a single byte. That amount to 256 indices, where 13 of them are reserved for the digits 0-9 and the characters '.', '-', and '*'. The remaining 243 are used to represent symbols of a given language. The digits 0-9 are reserved to express hyphenation rule levels, and the characters '.', '-', and '*' are reserved to express an incorrect hyphenation point, a missing hyphenation point, and a correct hyphenation point, respectively. To run *patgen*, a translation file is necessary. This file defines the values of certain language-specific parameters (in the first line) and enumerates the various forms in which language symbols may appear (all subsequent lines). In the first line, positions 1 and 2 are used to set the value of `lefthyphenmin`, and positions 3 and 4 are used to set the value of `righthyphenmin`. These values determine the minimum length of a string that may be generated by a hyphenation procedure. To set a single-digit value, leave the first position blank, i.e., place a space in position 1 and 3 for `lefthyphenmin` and `righthyphenmin`, respectively. Positions 5, 6 and 7 are used to define alternative values for the special characters '.', '-' and '*'.

1 2	3 4	5	6	7
lefthyphenmin	righthyphenmin	.	-	*

The following lines use a delimiter to enclose each 'letter' of the desired language alphabet, including its alternative representations. The first position of the line defines the delimiter, and each symbol of the language can occupy as many positions as necessary, as long as the reserved value for the delimiter is not used in the symbol's definition. The defining lines ends when the delimiter appears twice in a roll. Consider the following example for defining the letter 'e' in Portuguese: `XeXEXéXêXÉXÊX\'{e}X\^{e}X\'{E}X\^{E}XX`.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
delimiter	letter e	delimiter	letter É	delimiter	letter é	delimiter	letter ê	delimiter	letter Ê	delimiter	letter Ê	delimiter	TeX sequence for é	delimiter	TeX sequence for ê	delimiter	TeX sequence for É	delimiter	TeX sequence for Ê
X	e	X	E	X	é	X	ê	X	É	X	Ê	X	\'{e}	X	\^{e}	X	\'{E}	X	\^{E}

We have adopted `X` as the delimiter. This line represents the many ways in which the letter 'e' might be found: lowercase, uppercase, with or without acute or circumflex accents. Note that we have used the direct input

[5] Some examples of hyphenation dependency on characters far from the break point: *dem-o-crat* and *de-moc-ra-cy*; *as-pi-rin* and *aspir-ing*; *de-mon-stra-tive* and *dem-on-stra-tion*.

[6] *Glyph* is commonly used in linguistics, typography, and computer graphics to refer to a specific graphical representation of a character or symbol, which can be the entire symbol or a distinct visual element within it.

(using UTF-8 or other encoding that support the accented 'e' character) and also the compositional counterpart using the appropriate TEX control sequence that instructs TEX to place accent on the letter. In this definition, we have assumed that the many forms in which we may find the character e will be equivalent for hyphenation (pattern matching) purpose. As another example, see the next line which defines the character π (taken from Haralambous (2021)): `#p#P#\varpi ##`.

Patgen also needs a dictionary file, which is a list of pre-hyphenated words from which *Patgen* extracts patterns to create hyphenation rules. To ensure *Patgen*'s proper functionality, the translation file and the dictionary file must utilize the same encoding, even if it is a multi-byte encoding. The translation file describes how to handle byte sequences representing a glyph, and *Patgen* will work seamlessly when there are at most 243 symbols in the given language.

The syntax to run *patgen* is described in Listing 2

Listing 2: Syntax to run patgen.

```
patgen dictionary_file initial_pattern_set output_file translation_file
```

The dictionary file is a list of correctly hyphenated words, one per line; the initial pattern file is a set of hyphenation rules to be used as a starting point; the output file is the set of final rules created by *patgen*; and the translation file which maps the many forms each symbol in a language might appear in TEX documents. If it is desired to run *patgen* from scratch, starting from an empty set of rules, just use an empty file as the initial pattern set.

Patgen uses a few parameters along its executions:

hyph_start, hyph_finish : Two numbers between 1 and 9 (separated by a space), representing the desired pattern levels in the final set of rules. Odd pattern levels are hyphenating levels and even pattern levels are inhibiting levels. Higher level numbers prevail over lower ones, creating exceptions, exceptions over exceptions, and so on... `hyph_start` and `hyph_finish` specify the first and last levels, respectively, to be considered during the rule creation process.

pat_start, pat_finish : Patterns at each level are chosen in order of increasing pattern length (usually starting with length 2). This is controlled by the parameters `pat_start` and `pat_finish` specified at the beginning of each level. These are the minimum and maximum lengths of patterns we are interesting in. Their values range between 1 and 15.

good weight, bad weight, threshold : Each level of patterns is tested over all words in the dictionary. A pattern is only if it satisfies the following formula: $\alpha \times$ #good matches $- \beta \times$ #bad matches $\geq \eta$, where α is the `good weight`, β is the `bad weight` and η is the `threshold`.

4 TEX hyphenation rules for portuguese

Rezende (1987) was the first to create patterns for Portuguese hyphenation in TEX. The pattern set was last updated in 2015, incorporating contributions from Rezende and Almeida (2015), resulting in a total of 307 rules[7]. These rules effectively hyphenate the majority of Portuguese words.

However, in light of certain cases, we propose an analysis of the default rules to identify areas for improvement. By addressing specific issues and considering non-typical patterns, we aim to enhance hyphenation accuracy. The methodology used for this analysis is described in Section 5.

Out of the default TEX rules, 252 (82%) follow the pattern `1CV`, which represents the recurring CV syllables in Portuguese. As for the consonants, there are typically 18 considered (`b`, `c`, `ç`, `d`, `f`, `g`, `j`, `k`, `l`, `m`, `n`, `p`, `r`, `s`, `t`, `v`, `x`, and `z`) that can combine with 14 vowels (`a`, `e`, `i`, `o`, `u`, `á`, `â`, `ã`, `é`, `í`, `ó`, `ú`, `ê`, and `õ`), resulting in these CV patterns that indicate favorable hyphenation points before the consonants[8]. It is worth noting that there might be other rules or exceptions in the hyphenation patterns not covered by these default rules. To accommodate exceptions, Rezende (1987) proposes the following rules:

[7]The set of rules for Portuguese hyphenation is short when compared to rules in other languages. For example, English currently has 4938 rules, Russian has 7023 rules and German has 34011 rules.

[8]The complete list of these 252 patterns is given here: 1ba, 1be, 1bi, 1bo, 1bu, 1bá, 1bâ, 1bã, 1bé, 1bí, 1bó, 1bú, 1bê, 1bõ, 1ca, 1ce, 1ci, 1co, 1cu, 1cá, 1câ, 1cã, 1cé, 1cí, 1có, 1cú, 1cê, 1cõ, 1ça, 1çe, 1çi, 1ço, 1çu, 1çá, 1çâ, 1çã, 1çé, 1çí, 1çó, 1çú, 1çê, 1çõ, 1da, 1de, 1di, 1do, 1du, 1dá, 1dâ, 1dã, 1dé, 1dí, 1dó, 1dú, 1dê, 1dõ, 1fa, 1fe, 1fi, 1fo, 1fu, 1fá, 1fâ, 1fã, 1fé, 1fí, 1fó, 1fú, 1fê, 1fõ, 1ga, 1ge, 1gi, 1go, 1gu, 1gá, 1gâ, 1gã, 1gé, 1gí, 1gó, 1gú, 1gê, 1gõ, 1ja, 1je, 1ji, 1jo, 1ju, 1já, 1jâ, 1jã, 1jé, 1jí, 1jó, 1jú, 1jê, 1jõ, 1ka, 1ke, 1ki, 1ko, 1ku, 1ká, 1kâ, 1kã, 1ké, 1kí, 1kó, 1kú, 1kê, 1kõ, 1la, 1le, 1li, 1lo, 1lu, 1lá, 1lâ, 1lã, 1lé, 1lí, 1ló, 1lú, 1lê, 1lõ, 1ma, 1me, 1mi, 1mo, 1mu, 1má, 1mâ, 1mã, 1mé, 1mí, 1mó, 1mú, 1mê, 1mõ, 1na, 1ne, 1ni, 1no, 1nu, 1ná, 1nâ, 1nã, 1né, 1ní, 1nó, 1nú, 1nê, 1nõ, 1pa, 1pe, 1pi, 1po, 1pu, 1pá, 1pâ, 1pã, 1pé, 1pí, 1pó, 1pú, 1pê, 1põ, 1ra, 1re, 1ri, 1ro, 1ru, 1rá, 1râ, 1rã, 1ré, 1rí, 1ró, 1rú, 1rê, 1rõ, 1sa, 1se, 1si, 1so, 1su, 1sá, 1sâ, 1sã, 1sé, 1sí, 1só, 1sú, 1sê, 1sõ, 1ta, 1te, 1ti, 1to, 1tu, 1tá, 1tâ, 1tã, 1té, 1tí, 1tó, 1tú, 1tê, 1tõ, 1va, 1ve, 1vi, 1vo, 1vu, 1vá, 1vâ, 1vã, 1vé, 1ví, 1vó, 1vú, 1vê, 1võ, 1xa, 1xe, 1xi, 1xo, 1xu, 1xá, 1xâ, 1xã, 1xé, 1xí, 1xó, 1xú, 1xê, 1xõ, 1za, 1ze, 1zi, 1zo, 1zu, 1zá, 1zâ, 1zã, 1zé, 1zí, 1zó, 1zú, 1zê, 1zõ.

- 20 rules created for cases involving consonants `b`, `c`, `d`, `f`, `g`, `k`, `p`, `t`, `v`, or `w` followed by `l` or `r`[9];
- 3 rules for `c`, `l` or `n` followed by `h`[10];
- 23 distinct patterns were introduced to indicate hyphenation points between vowels or between `c`'s, `r`'s, and `s`'s[11];
- 8 patterns adhere to the `1[gq]u4V` pattern, signaling beneficial hyphenation points before `g` or `q`, followed by a sequence of `u` and a vowel, with an inhibiting point between the `u` and the subsequent vowel[12];
- 1 pattern represented as `1-`, denoting that a hyphen indeed serves as a beneficial hyphenation point.

5 Methodology

To assess the performance of each set of rules, we require a collection of correctly hyphenated Portuguese words. We employed word frequency data to ensure we selected representative, commonly used words, avoiding excessively rare ones. Our comprehensive word list was curated using online dictionaries, Portuguese corpora, and verified against Portuguese grammars to validate hyphenation rules. We also performed manual corrections where necessary. We should not expect an entire corpus to follow a set of conventions for a written language nor uniformity among a variety of corpora, especially those harvested from the Internet. "It is notoriously difficult to prescribe rules governing the use of a written language; it is even more difficult to get people to 'follow the rules.' This is in large part due to the nature of written language, in which the conventions are not always in line with actual usage and are subject to frequent change" (Palmer 2010, p. 14).

5.1 Data collection

Initially, we selected CETENFolha as the source corpus, but this approach presented two issues: many words were still missing, and the corpus was based on text from 1994, prior to the implementation of the Orthographic Agreement. The Orthographic Agreement of the Portuguese Language was conducted in 1990, and its transition period started in 2009, becoming mandatory in 2016. However, even after the agreement, some words continue to have different spellings in the participating countries. To address these idiosyncrasies, we aim to develop hyphenation rules that accommodate variations in spelling. For instance, the word *reception* might be written as *receção* (Portugal) or *recepção* (Brazil); the word *action* might be written as *acção* (Portugal) or *ação* (Brazil); and the word *project* might be written as *projecto* (Portugal) or *projeto* (Brazil).

Due to the peculiarities mentioned, we made the decision to integrate the word list from Palavras NET, and subsequently, we augmented it with words sourced from the Portuguese Wikipedia dump. From the Wikipedia data, we narrowed down the selection to a subset of 50 721 words, accounting for 95% of occurrences in the corpus. This threshold was set to filter out typos and infrequent words. The initial word list became quite extensive, comprising 419 578 words. However, we refined the list by retaining only those words for which we could find hyphenation data in at least one of the following online dictionaries: Michaelis, Priberam, Wikcionário, Aulete, Portal da Língua Portuguesa, and Dicio. Consequently, this curation process resulted in a final dictionary containing 85 638 words.

Table 1 presents the number of distinct words with a given number of hyphenations found in the dictionaries. Additionally, it displays the number of words that received the highest agreed-upon hyphenation. For instance, the word *como* was consistently hyphenated as *co-mo* in all six dictionaries, therefore, it contributes to both the counts in the column representing six hyphenations (first column). On the other hand, the word *sua* received different hyphenations: *su-a* in four dictionaries, *sua* in one dictionary, and one dictionary did not provide any hyphenation result. This leads to an increment in the first line, second column (since hyphenations were found in five dictionaries) and the second line, third column (since the most frequent hyphenation was found in four dictionaries). It is evident now that the number of identical hyphenations can exceed the total number of hyphenations found, especially for smaller values.

In summary, our data compilation involved *CETENFolha*, *Palavras NET*, and *Wikipedia*, and hyphenated the words using the original TEX hyphenation rules. Subsequently, we compared the hyphenated results with those obtained from the dictionary. By conducting a thorough performance appraisal, we systematized the errors and delved into the relevant literature to identify potential rules that could complement and enhance the existing hyphenation system.

[9] The 20 additional rules are: 1b2l, 1b2r, 1c2l, 1c2r, 1d2l, 1d2r, 1f2l, 1f2r, 1g2l, 1g2r, 1k2l, 1k2r, 1p2l, 1p2r, 1t2l, 1t2r, 1v2l, 1v2r, 1w2l, 1w2r.

[10] Those 3 rules are: 1c2h, 1l2h, 1n2h.

[11] Those patterns are: a3a, a3e, a3o, c3c, e3a, e3e, e3o, i3a, i3e, i3i, i3o, i3â, i3ê, i3ô, o3a, o3e, o3o, r3r, s3s, u3a, u3e, u3o, u3u.

[12] They are the following: 1gu4a, 1gu4e, 1gu4i, 1gu4o, 1qu4a, 1qu4e, 1qu4i, 1qu4o.

Table 1: Considering the six dictionaries used, the first line of this table presents the number of words that a given number of hyphenations were found in the dictionaries. The second line presents the number of words that have a given number of hyphenation as its most frequent form found in the dictionaries.

hyphenations	6	5	4	3	2	1	0
hyphenations found	16 036	15 820	10 325	7 783	8 520	27 840	333 253
same hyphenations	15 842	15 642	10 299	7 745	8 435	28 368	n/a

6 Writing systems

There is a diverse collection of writing systems, categorized into logographic, syllabic, and alphabetic systems. Although distinct, these systems can be built on an interplay of these categories (Coulmas 2003; Palmer 2010). The principles guiding an alphabetic or phonemic writing system vary significantly based on the language and its history. Here are some key points worth highlighting:

1) The most common principle involves representing the sounds of a language with written symbols, using one or a combination of symbols to represent each sound (e.g., the orthographic system of Portuguese uses an alphabet of 27 letters to represent different sounds of the language).

2) Etymology also plays a crucial role, where the spelling of a word reflects its origins and historical development (e.g., the French language often reflects the Latin roots of words in its spelling).

3) Morphology serves as a guide in structuring many languages, using specific letters or symbols to indicate word endings, prefixes, or suffixes (e.g., Russian uses different forms of the Cyrillic alphabet to indicate gender and case in its nouns).

4) The evolution of a language over time also contributes to its written form (e.g., the spelling of many English words has changed over time to reflect changes in pronunciation).

Orthographic systems for languages with logographic writing systems, such as Chinese ideograms, cuneiform writing, and Egyptian hieroglyphs, differ significantly from alphabetic writing systems. In logographic systems, individual symbols or characters represent entire words or ideas, rather than phonemes or sounds. As a result, the principles guiding their orthographic systems are based more on semantic and visual principles than on phonemic principles. Hyphenation is a feature of alphabetic scripts where words are composed of letters, and spaces or hyphens are used to separate words, syllables, or parts of words. Logographic writing systems, such as the Chinese, function differently and do not employ hyphenation (Honorof and Feldman 2006).

7 Portuguese spelling system

Portuguese employs an alphabetical writing system, which means its spelling is guided by phonological principles (Cagliari 2015). The correlation between spelling and pronunciation influences word hyphenation, as words are divided into syllables based on the phonemic system. It is important to note that different languages follow diverse principles for word division. For example, English is primarily guided by morphological principles, evident in words like *walk-ing*, *un-happy*, *work-s*, and *ear-ly*. Additionally, other factors also impact word hyphenation in English, such as the distinction between long and short vowels, which function within the context of open or closed syllable, respectively; and the presence of doubled consonants and digraphs (Lin 2011; Yavas 2020). Although each language has its driving principles for the hyphenation process, multiple factors come into play, leading to various solutions in certain scenarios. For instance, in Portuguese, the phonological principle would lead to *hipe-rativo*, while the morphological principle would lead to *hiper-ativo*. Both approaches seem valid and are indeed found in online dictionaries[13]. Some rather rare words, such as *hiperalgesia*, may not be subject to morphological influences. Since it has a low frequency of occurrence, the individual may not be aware of its morphological components and therefore hyphenate it as *hipe-ralgesia*. The form *hiper-algesia* could also be accepted, emphasizing its morphological constituents. Furthermore, there exist numerous exceptions that may be categorized into rules.

Merely stating that an orthographic system is guided by phonological issues does not necessarily mean that its hyphenation rules directly mirror the phonetic counterpart. This is notably apparent in Portuguese, where a strict one-to-one correspondence between letters and sounds is not always observed. The orthographic system operates according to its specific rules. For example, consider consonant clusters that create a single sound (digraphs) in words like *a<u>ch</u>ado*, *i<u>lh</u>a*, *su<u>sh</u>i*, *ca<u>rr</u>o*, and *mas<u>s</u>a*. While these digraphs are pronounced as one sound within a single syllable, their representation in writing determines how they are divided. Specifically,

[13]Looking at words with prefixes, it seems that even on those words, the phonological approach was predominant on hyphenation, but we cannot tell if it is a byproduct of automatic hyphenation that might be used to hyphenate words on online dictionaries.

different consonants within a digraph must remain together, whereas identical consonants are separated. As a result, we observe hyphenations like *a-cha-do*, *i-lha*, and *su-shi*, but *car-ro* and *mas-sa*.

In Portuguese, hyphenations are allowed on syllables boundaries and, in general, follow phonological principles. According to the Grammar (Cunha and Cintra 2016; Bergström and Reis 2011; Cegalla 2020), some rules might still apply:

Non-Splitting Rules

1. diphthong or triphthong should not be split (e.g. *mui-to*, *Pa-ra-guai*);
2. the sequences *ia*, *ie*, *io*, *oa*, *ua*, *ue* and *uo*, when in final unstressed position, should not be split (e.g. *gló-ria*, *vi-tó-ria*, *cá-rie*, *es-pé-cie*, *Má-rio*, *má-goa*, *ré-gua*, *tê-nue*, *con-tí-guo*, *am-bí-guo*);
3. consonant clusters starting a syllable should not be split (e.g. *pneu-má-ti-co*, *psi-có-lo-go*, *mne-mô-ni-co*);
4. the digraphs *ch*, *lh*, *nh* should not be split (e.g. *ra-char*, *a-bro-lhos*, *ma-nhã*;
5. bigrams like *gu* and *qu* whose vowel *u* is not pronounced are never separated from the vowel or diphthong that follows it (e.g. *U-ru-guai*, *pe-que*);
6. since they are digraphs, a vowel and its following nasalization marker (a graphic nasal consonant) should not be split (e.g. *am-bição*, *man-cha*);
7. decreasing diphthongs should not be split (e.g. *ai-ro-so*, *ca-dei-ra*, *o-ra-ção*);
8. rising diphthong should not be split (e.g. *a-bai-xo*; *cau-te-la*, *pa-péis*, *cha-péu*, *pre-fei-to*, *reu-nir*, *noi-te*; *ca-la-bou-ço*, *as-te-rói-de*; *re-tri-bui*);
9. disyllables whose syllable has a single vowel should not be split (e.g. *ato*, *rua*, *ódio*, *unha*);
10. words with more than two syllables, when divided, cannot isolate a syllable composed of a single vowel (e.g. *agos-to*, *la-goa*, *ida-de*);

Splitting Rules

11. hiatus vowels and those vowel sequences where each vowel belongs to a different syllable should be split (e.g. *sa-ú-de*, *ra-i-nha*, *do-er*, *vo-os*), the same procedure is used splitting diphthongs in different syllables (e.g. *cai-ais*) or diphthong and vowel in different syllables (e.g. *en-sai-os*);
12. consonant sequences, when in different syllables, should be split (e.g. *af-ta*, *ab-di-car*, *res-ci-são*, *ab-so-lu-to*);
13. the following consonant digraphs should be split: *rr*, *ss*, *mm*, *nn*, *sc*, *sç* and *xc* (e.g. *ter-ra*, *pro-fes-sor*, *co-mum-men-te*, *con-nos-co*, *des-cer*, *cres-ça*, *ex-ce-der*).

Rules 9 and 10 are primarily aimed at ensuring proper readability of the text, aligning with TEX approach to deal with widows and orphans. As mentioned in Section 1, the variables `lefthyphenmin` and `righthyphenmin` control the minimum length for fragments of hyphenated words. When those variables are set to values greater than one, Rules 9 and 10 become fiddling rules in TEX hyphenation, and they could be disregarded. Notwithstanding, the full hyphenation of words is useful, particularly in text-to-speech applications (Libossek and Schiel 2000; Trogkanis and Elkan 2010).

Additionally, it is advisable to refrain from splitting disyllables consisting of four letters (e.g., *para*, *como*, *cede*). This considerations also lead to more aesthetically pleasing and intelligible text, and TEX's control of isolated fragments, through the variables `\lefthyphenmin` and `\righthyphenmin`, already address this issue.

In some situations a hyphen should be repeated at the start of the following line. They are:

1. cases where compound words using a hyphen are split across lines (e.g., *couve-/-flor*, *ex-/-presidente*); and
2. cases where splitting a pronoun could result in a different meaning (e.g., *prazer de ver-/-me*[14]).

Systematizing the rules that guide syllable boundaries and hyphenation in Portuguese is a fundamental step to understand and improve the TEX hyphenation rules. To ensure the accuracy of hyphenation rules and effectively compare their results, a hyphenation spelling dictionary serves as an indispensable reference. This dictionary provides a comprehensive listing of correctly hyphenated words, enabling the computation of the correctness of each set of hyphenation rules. By consulting the hyphenation spelling dictionary, one can verify whether the hyphenation patterns generated by a set of rules align with established standards. This process involves analyzing how well the hyphenation rules adhere to the accepted conventions of the language, ensuring that they effectively segment words without compromising readability or consistency. Moreover, by comparing the results obtained from different sets of hyphenation rules against the entries in the spelling dictionary, we can assess the efficacy and reliability of each approach. In section 8, we will define the specifications of this reference dictionary, outlining its role in subsequent rule updates and rules generation.

[14]Possible conveyed meanings: *pleasure in seeing me* or *worm's pleasure*.

8 Creating the dictionary

A bash script, named `gethyphenations.sh`, was developed to crawl six online dictionaries and extract word hyphenations whenever available. The process involves launching parallel threads for each word in the list, querying a specific online dictionary, and retrieving the corresponding hyphenation. For this purpose, individual scripts were created for each dictionary: `getmichaelishyphenation.sh` for Michaelis, `getpriberamhyphenation.sh` for Priberam, `getwiktionaryhyphenation.sh` for Wikcionário, `getauletehyphenation.sh` for Aulete, `getportalhyphenati` for Portal da Língua Portuguesa, and `getdiciohyphenation.sh` for Dicio. This systematic approach ensures efficient and accurate extraction of hyphenations from various online sources. The resulting data is stored in a CSV[15] file named `hyphenations.csv`.

8.1 Establishing a gold standard

To establish our gold standard for hyphenation, we will compare the hyphenations offered by all six dictionaries. First, we will determine the total count of unique hyphenations across these dictionaries using the script `hyphenationagreements.sh`. It's worth noting that certain dictionaries annotate words with a colon (':') to indicate potential variations in pronunciation or production. Particularly, 'apparent proparoxytones' might appear marked by this colon sign, indicating the possibility of a diphthongs or hiatuses in the word ending. In Portuguese, words ending in *ea*, *eo*, *ia*, *ie*, *io*, *ua*, *ue*, or *uo*, with the stressed syllable before these endings, can be considered either paroxytones ending in a rising diphthong or proparoxytones ending in a hiatus. For instance, words like *his-tó-ri:a* and *on-du-la-tó-ri:o* are examples of words marked as apparent proparoxytones, suggesting they could be regarded as paroxytones (*his-tó-ria* and *on-du-la-tó-rio*) or proparoxytones (*his-tó-ri-a* and *on-du-la-tó-ri-o*). The script `hyphenationagreements.sh` offers four potential approaches to handle these apparent proparoxytones: retaining the hyphenations with the colon mark to indicate an apparent proparoxytone; selecting only the paroxytone or only the proparoxytone counterpart; or counting both paroxytones and proparoxytones. Given the lack of consensus in the literature, our initial approach involves counting both hyphenations, and we will detail below how we resolve this issue. In our hyphenation dictionary, we have identified 1848 words marked as apparent proparoxytones.

As our initial hyphenation reference, we will consider those cases where all six dictionaries are in agreement. This comprehensive list comprises 15 842 words. Their corresponding hyphenations are meticulously compiled and stored in the file `hyphenations6.dic`. This curated list serves as a benchmark for evaluating the accuracy and effectiveness of hyphenation rules generated through computational methods. It will establish a reliable baseline for our hyphenation analysis, enabling us to assess and refine our methodologies effectively.

In case of disagreement, we must exercise caution and examine each instance. We might be tempted to choose the majority vote approach, simply selecting those hyphenations that have the most supporters, but it is unclear how the hyphenations on those online dictionaries were curated; many might use algorithmic approaches, leading to potential flaws shared among them. For example, 5 dictionaries use the hyphenation *quart-zo* while the dictionary Aulete (originally from Portugal) favors the hyphenation *quar-tzo*, which can be explained by the lack of an epenthetic vowel in European Portuguese[16] (Mateus n.d.). Similarly, *ap-nei-a* or *a-pnei-a* and *disp-nei-a* or *dis-pnei-a* should be accepted considering, since there could be an epenthetic vowel (Brazilian Portuguese) or not (European Portuguese). The same happens with *hiperalgesia*, already mentioned in Section 7. Among other examples, we might also highlight *neerlandês*, which appears hyphenated as *ne-er-lan-dês* in five dictionaries and as *neer-lan-dês* in one. If we consider its pronunciation, we would prefer the former hyphenation, but the first one conforms better to regular written syllables in Portuguese. Some strings might accept more than one hyphenation, since they represent different words. For example, *sub-li-nha* and *su-bli-nha*, where the former is the substantive *underline* and the latter is a flexed form of the verb *to underline*.

The procedure to create the second list of hyphenated words involved several steps: 1) select those hyphenations agreed by five dictionaries, that have no alternative hyphenation proposal (5 to 0 votes); 2) from those with alternative proposals, exam whether they constituted apparent proparoxytones (these cases could arise from instances with 5 to 1 votes or 6 to 1 votes, as explained earlier), a) if the word had a diacritic, select the paroxytone option (e.g. *po-lí-cia* instead of *po-lí-ci-a* and *pe-rí-neo* instead of *pe-rí-ne-o*), b) if not, choose the proparoxytone option (e.g. *au-to-ri-a* instead of *au-to-ri-a* and *mar-ce-nei-ro* instead of *mar-ce-neiro*)[17]. 3) if the word did not fall into the apparent proparoxytone category, select the hyphenation with five votes. This procedure was implemented in the script `get51.sh`. Only a few hyphenations required manual correction, including the following that were added: *quar-tzo*, *quar-tzi-to*, *su-bli-nha*, *a-pnei-a*, *dis-pnei-a*, *hi-per-al-ge-si-a*,

[15] Comma-separated values (CSV) is a text file format defined in RFC 4180. In CSV, commas are used to separate values, and each line in the file represents a new record. Records consist of fields separated by delimiters, typically a single comma.

[16] The online dictionary Aulete has its origins on the printed version known as *Dicionário Caldas Aulete*. It is a dictionary from Portugal, reflecting the European Portuguese dialect.

[17] We just have to make sure we are not selecting those cases where there is a sequence of the form `[gq]u-[aeio]` (using regex notation), as in *a-pa-zi-gu-ar* and *en-xa-gu-ar*.

and *neer-lan-dês*. The final list, containing 15 642 additional hyphenations, is stored in `hyphenations5.dic`, bringing the total number of hyphenations across both lists to 31 484.

For those hyphenations with four votes, we adopted a similar approach to the one used previously. Many words garnered only four votes, indicating a lack of alternative hyphenations provided by any dictionary, and they were thus initially hyphenated as proparoxytones. These instances were typically distinguished by diacritics in the third-to-last syllable and a *V*-*V* sequence at the end. To address this, we converted these words into paroxytones by eliminating the final hyphen. In cases where alternative hyphenations were available, we assessed the presence of diacritics. If diacritics were detected, we chose the paroxytone option; otherwise, we defaulted to the proparoxytone option. In instances with multiple alternative options, we selected the option with the highest frequency of votes. This procedure is implemented in the script `get4.sh`, and the resulting list has 10 299 words.

After completing the aforementioned procedures to create our golden hyphenation dictionary, manual corrections may still be necessary to ensure its accuracy and reliability. These manual adjustments, which encompass the corrections described earlier, are documented in the file `replacements.txt`. Utilizing the script `manualcorrections.sh`, we meticulously apply these corrections to refine our hyphenation dictionary and enhance its effectiveness.

9 Results for the default rules

In this section we will apprise the performance of default TEX hyphenation rules on the list of hyphenated words, which creation was described in Section 8. Table 2 summaries the results, showing the number of words correctly hyphenated words by the default rules, the number of words incorrectly hyphenated and the number of words where a hyphenation point was missed.

Table 2: Results of TEX default hyphenation rules on `hyphenations6`, `hyphenations5` and `hyphenations4` dictionaries.

word list	# correct	# incorrect	# missing	# entries
hyphenations6	15537	30	277	15842
hyphenations5	13981	1146	601	15642
hyphenations4	9001	883	518	10299
total	38519	2059	1396	41783

Each word in the list is hyphenated using the set of default rules using the `gohyphen` script (the `gohyphenfull` script is used to see in the interplay among all rules that apply to a given word) (Gundlach 2021). The hyphenation resulted from the given rules is compared to the hyphenation in our dictionary, so we may check if there were erros (incorrect or missing hyphenation points).

10 Updating the rules

In this section we will analyse the results of the default rules on our dictionaries, find patterns, and gradually incorporate new rules to the set, appraising the result of each one, to avoid causing more damage than good, create exception rules when necessary, and, if necessary, discard rules. We will stick to the *patgen* notation to express a wrong, a missing, and a correct hyphenation point ('.', '-', and '*', respectively).

For the list comprised in `hyphenations6`, below is the complete list of 30 words that were incorrectly hyphenated:

*p.si*co*lo*gi*a*
*p.si*co*ló*gi*co*
*p.si*qui*a*tra*
*p.si*ca*na*lis*ta*
*p.si*có*lo*go*
*p.si*qui-á*tri*co*
*t.che*co*
*p.neu*mo*ni*a*
*p.sí*qui*co*
*p.si*co*pa*ta*
*p.si*co*se*
*p.si*co*te*ra*pi*a*
*p.si*ca*na*lí*ti*co*
*p.si*co*te*ra*peu*ta*
*p.si*có*ti*co*
*g.no*mo*
*p.si*co*lo*gis*mo*
*p.si*quis*mo*
*p.si*co*mo*tor*
*t.me*se*
*p.neu*má*ti*ca*
*p.neu*má*ti*co*
*p.si*cos*so*má*ti*co*
*p.si*co*tró*pi*co*
*g.nos*ti*cis*mo*
*g.nós*ti*co*
*p.si*co*mo*tri*ci*da*de*
*p.si*co*pa*to*ló*gi*co*
*p.to*se*
*su.b-lu*nar*

In each instance, an erroneous hyphenation point was placed at the beginning of a word, indicating a potential issue in recognizing certain prefixing morphemes in the language. Among those, 20 occurred in

words containing the *psi* morpheme, where the algorithm erroneously hyphenated between the initial *p* and the following *s*. Additionally, 3 cases involved the starting *gno* sequence, while the remaining errors were observed in the sequences *pneu* (3), *tch* (1), *tme* (1), *pto* (1), and *sub* (1).

Those erroneous hyphenation points might be corrected by the introduction of a few rules: `.p2si, .p2sí` (see rule 10), `.g2no, .g2nó` (see rule 1), `1p2neu` (see rule 3), `t2c` (see rule 2), `.t2m` (see rule 4), `.p2t` (see rule 5) and `su2b3r`, `su2b3l` (see rule 11). They are deeply discussed in Section 13.

The first 20 examples of missing hyphenations are displayed in the following list:

*a-in*da*	*ca-ir*	*a*tra-ir*
pa-ís	*pa*ra-í*ba*	*subs*ti*tu-ir*
*re*gi-ão*	*pre*ju-í*zo*	*in*clu-ir*
*sa-ú*de*	*cons*tru-ir*	*ga-ú*cho*
a-í	*ve-í*cu*lo*	*ra-i*nha*
sa-ir	*co*ca-í*na*	*re*li*gi-ão*
da-í	*pa-ul*	*...*

It is a long list with 277 entries, therefore it is unavoidable to analyse it through clusters of certain patterns. It is presented below the counts of immediate context in which those missing hyphenations were found:

38 u-i	*6 i-ú*	*2 o-á*	*1 s-q*	*1 e-ê*
22 e-í	*5 r-q*	*2 i-é*	*1 o-w*	*1 e-é*
20 a-i	*5 e-ó*	*2 e-ú*	*1 o-l*	*1 e-â*
19 a-í	*5 e-i*	*2 é-o*	*1 o-i*	*1 b-l*
17 u-í	*5 a-q*	*2 e-l*	*1 o-g*	*1 a-u*
14 í-a	*4 a-v*	*2 e-c*	*1 o-é*	*1 a-t*
12 i-á	*3 o-ó*	*2 e-á*	*1 o-b*	*1 a-r*
11 i-ó	*3 i-u*	*2 a-é*	*1 i-v*	*1 a-p*
11 a-ú	*3 e-q*	*1 u-q*	*1 i-q*	*1 a-l*
10 i-ã	*3 e-ã*	*1 ú-o*	*1 í-o*	
9 o-í	*3 a-ó*	*1 u-l*	*1 i-í*	
6 o-q	*2 u-á*	*1 u-ã*	*1 i-c*	

On the top of the list we see the missing hyphen in *u-i*, accounting for 38 cases. Among those, 34 were originated from the pattern `u-ir` in the end of word and 13 from `a-ir` in the end of word. To fix them, We will include the rules `u1ir.` and `a1ir.` (see rule 15). The cases that are not covered by this rule are: *tu-im*, *je*su-i*tis*mo*, *ma*lau-i*a*no*, and *con*tri*bu-in*te*. On rule 21 we see how to deal with them.

The default rules already include 17 separating rules for vowel sequences:

`a3a`	`e3e`	`i3i`	`i3ô`	`u3a`
`a3e`	`e3o`	`i3o`	`o3a`	`u3e`
`a3o`	`i3a`	`i3â`	`o3e`	`u3o`
`e3a`	`i3e`	`i3ê`	`o3o`	`u3u`

but there are just 3 rules to hyphenate between vowels when one has a diacritic: `i3â`, `i3ê`, `i3ô`. We could then add more 29 rules to account for the missing hyphenations between vowels with diacritics and also the missing rule `i1u`:

`a1é`	`e1ã`	`é1o`	`i1ó`	`o1í`
`a1í`	`e1é`	`i1á`	`i1u`	`o1ó`
`a1ó`	`e1ê`	`i1ã`	`i1ú`	`u1á`
`a1ú`	`e1í`	`í1a`	`í1o`	`u1ã`
`e1á`	`e1ó`	`i1é`	`o1á`	`u1í`
`e1â`	`e1ú`	`i1í`	`o1é`	`ú1o`

See more about this on rule 22. These rules will require exceptions for sequences with *[qg]uV'* (where we used *V'* to represent a vowel with a diacritic). See more on rule 22.

From the list of missing hyphenations above, we find another recurring pattern, a missing hyphen preceding *q*:

a-q	*i-q*	*r-q*	*u-q*
e-q	*o-q*	*s-q*	

That issue might be easily solved by adding a hyphenation rule `1qu`. The *q* is always followed by a sequence *uV*, and there are already four hyphenation rules involving this sort of sequence in the default rules (`1qu4a`, `1qu4e`, `1qu4i` and `1qu4o`)[18]. The only valid exceptions[19] found in Wikipedia corpus are the words *far-quhar* and *qu-bit*. We could then add the rule `1qu` along with the exception rules for those hyphenation rules between *u* and the following vowel (see rule 22): `1qu2á`, `1qu2ã`, `1qu2â`, `1qu4é`, `1qu2ê`, and `1qu2í` (see rule 22). Note that `1qu2ó` was not added since we will not use a rule to hyphenate between these vowels.

The missing hyphens in *a-v* and *i-v* might be easily solved by introducing the rule `1vô` (see rule 23) which complements the rule `1vo` already included in the default set of rules. Similarly, the missing hyphens *e-l*, *u-l*, *o-l*, and *a-l* are corrected by the introduction of the rule `1lô` rule (see rule 23), complementing the already included rule `1lo`. In the same way, the rules `1cô`, `1gô`, `1bô`, `1tô`, `1rô`, and `1pô` (see rule 23) should be added to fix the missing hyphens *e-c*, *i-c*, *o-g*, *o-b*, *a-t*, *a-r*, and *a-p*, respectively. The missing hyphen *b-l* happened in *su.b-lu*nar*, which also has a wrong hyphenation. Those errors were caused by the default rule `1b2l`. The introduction of rule `su2b3l` will help solve this matter (see details in rule 11).

The remaining missing hyphen is *o-w*, which comes from the foreign *kilo-watt*. Those cases of words borrowed from other languages are dealt in the topic Foreignness.

After introducing the rules: `.p2si`, `.p2sí`, `.g2no`, `.g2nó`, `t2c`, `1p2neu`, `.t2m`, `.p2t`, `su2b3r`, and `su2b3l`, the number of incorrect hyphenations decreased to 3, and the number of missing hyphenations decreased to 278. At first sight, we expected the number of incorrect hyphenations to drop to zero. However, the introduction of rule `su2b3l` introduced hyphenation errors in *su-b.li*me*, *su-b.li*ma*ção* and *su-b.li*ma*do*.

Listing 3: Hyphenation of the word *sublime* after the indroduction of rule `su2b3l`. This rule introduces a wrong hyphenation point, demanding an exception rule to fix it.

```
    .   s   u   b   l   i   m   e   .
      1   0   0   |   |   |   |   |     1su
      0   0   2   3   0   |   |   |     su2b3l
      |   |   1   2   0   |   |   |     1b2l
      |   |   |   1   0   0   |   |     1li
      |   |   |   |   |   1   0   0     1me
max:  1   0   2   3   0   1   0   0
final: s   u   b - l   i - m   e
```

By inserting the rule `.su3b4li`, we fixed the hyphenation errors caused by the introduction of the rule `su2b3l`, as illustrated in Listing 3. This adjustment resulted in zero hyphenation errors and 276 missing hyphens.

The next step is to introduce the rules to address the missing hyphenation points. These rules are: `a1ir.`, `u1ir.`, `1qu`, `1vô`, `1lô`, `1cô`, `1gô`, `1bô`, `1tô`, `1rô`, `1pô`, `a1é`, `a1í`, `a1ó`, `a1ú`, `e1á`, `e1â`, `e1ã`, `e1é`, `e1ê`, `e1í`, `e1ó`, `é1o`, `e1ú`, `i1á`, `i1ã`, `í1a`, `i1é`, `i1í`, `i1ó`, `í1o`, `i1u`, `i1ú`, `o1á`, `o1é`, `o1í`, `o1ó`, `u1á`, `u1ã`, `u1í`, `ú1o`. After inserting these rules, the number of incorrect hyphenations increases to 18 due to the inclusion of hyphens in *qu.í*, *gu.í* and *qu.á*. To overcome these errors, we need to add the exception rules: `1qu2á`, `1qu2í`, and `1gu2í`. See more on this on rule 22.

We have reached a point where there is no wrong hyphenation but still there are 19 missing hyphenation. Here is the list of the missing ones:

*a-in*da*	*re-in*te*gra*ção*	*re-in*ci*dir*	*tu-im*	*re-im*pri*mir*
*ra-i*nha*	*cam*pa-i*nha*	*pa-in*ço*	*je*su-i*tis*mo*	
*con*tri*bu-in*te*	*re-in*te*grar*	*pi*xa-im*	*ben*jo-im*	
ra-iz	*ta-i*nha*	*re-im*pres*são*	*ma*lau-i*a*no*	

Observing the instances above, 17 out of 19 involve a missing hyphen before the vowel *i*. The following rules will be used to solve those issues: `a1ind`, `a1i1nh` (see rule 16), `e1imp` (see rule 17), `e1inc`, `e1inf`, `e1ing`, `e1ins`, `e1int`, `e1inv`, (see rule 18), and `u1iz.`, `a1iz.` (see rule 19).

After incorporating these rules, the number of missing hyphenations decrease to only 8:

[18]The default rules uses level three to hyphenate between those vowels: `u3a`, `u3e`, `u3i`, and `u3o`. Apparently, they could have used a lower degree rule and, consequently, a lower degree rule to discourage the hyphen in sequences with *q* or *q*

[19]The other exceptions we are not considering include typos and proper names (names of places, for example *Urquhart*, *Qurgonteppa* , *Al-Qusayr*, etc.).

pa-ul	*pa-in*ço*	*tu-im*	*ben*jo-im*
*qui*lo-watt*	*pi*xa-im*	*je*su-i*tis*mo*	*ma*lau-i*a*no*

Since these remaining cases comprises only a few rare words, we now proceed to integrate the dictionary `hyphenations5` and evaluate the results of the current rules against this set. In this new dataset, the number of incorrect hyphenations increases to 1121, with the majority stemming from apparent proparoxytones. These cases can be easily rectified by avoiding the final hyphen which leaves the vowels *a*, *e* or *o* alone in the last syllable. To address this, we introduce rules `4a.`, `4e.`, `4o.` (see rule 24). While a more comprehensive rule set could be developed to account for all scenarios[20], we opted for these more concise rules despite their limitations. Note that, this set of three rules might be easily suppressed if someone intend to get the final vowel hyphenated. The addition of these rules reduces the number of wrong hyphenations to 35; however, it increases the number of missing hyphenations from 133 to 900. Among these, 770 additional missing hyphenations occur before the final vowel of a word. Considering the smaller number of cases, the perceived severity (a missing hyphen is deemed less problematic than an incorrect one), and the position of the word's end, we have chosen to retain these rules.

The resulting errors are shown in the following list:

*sa*gu.ão*	*c.za*ris*ta*	*p.so*rí*a*se*	*núp*ci.as*	*flâ*mu.la*
*p.seu-dô*ni*mo*	*bre.ch*ti*a*no*	*t.za*ris*ta*	*su.b-li*nha*	*a-p.nei-a*
*s.ta*li*nis*mo*	*c.za*ris*mo*	*g.nais*se*	*a*rá*bi.as*	*c.ni*dá*rio*
*a.b-rup*to*	*su.b-li*te*ra*tu*ra*	*m.ne-mô*ni*co*	*quar-t.zo*	*dis-p.nei-a*
*su.b-li*mi*nar*	*ca*ra*min*gu.ás*	*p.so*as*	*e*fe*mé*ri.de*	*gli*có*li.se*
*quar-t.zi*to*	*d.ze*ta*	*su.b-lin*gual*	*cri*sân*te.mo*	*hi*pe.r-al*ge*si-a*
*c.za*ri*na*	*m.ne-mô*ni*ca*	*es*tô*ma.go*	*e*xé*qui.as*	*ne.er*lan*dês*

From these, we have proposed the rules: `1gu2á`, `1gu2ã`, `1qu2ã` (see rule 22), `.m2n` (see rule 6), `c2za` (see rule 7), `.s2` (see rule 8), and `1p2seu1d` (see rule 13).

The inclusion of these rules leave us with 21 incorrect hyphenations and 899 missing ones. The list of wrong hyphenations follows below:

*a.b-rup*to*	*d.ze*ta*	*su.b-lin*gual*	*e*xé*qui.as*	*ne.er*lan*dês*
*su.b-li*mi*nar*	*p.so*rí*a*se*	*núp*ci.as*	*a-p.nei-a*	
*quar-t.zi*to*	*t.za*ris*ta*	*su.b-li*nha*	*c.ni*dá*rio*	
*bre.ch*ti*a*no*	*g.nais*se*	*a*rá*bi.as*	*dis-p.nei-a*	
*su.b-li*te*ra*tu*ra*	*p.so*as*	*quar-t.zo*	*hi*pe.r-al*ge*si-a*	

Most of these cases occur in rare words or loanword. We decided not to address them. Analyzing the missing hyphenation points, as mentioned earlier, most result from the rules (`4a.`, `4e.`, `4o.`) we proposed to handle apparent proparoxytones. This leave us with 125 missing hyphenations to address. We observe that a few additional rules with diacritic counterpart need to be introduced: `1dô`, `1fô`, `1mô`, `1nô`, `1sô`, `1zô` (see rule 23), `u1é`, `1gu2é`, `1qu2é` (see rule 22). After incorporating these rules, we now have 21 incorrect and 856 missing hyphenations, 86 of which do not involve the final vowel.

To account for a few more missing cases, we propose the inclusion of the following rules: `tu1i`, `bu1i`, `nu1i`, `o1in`, `u1in`, `su1i`, `í1e`, `ju1i`, `fu1i`, `du1i`, `do1im`, `au1i`, `u1i1ç` (see rule 21). With these additions, the number of missing hyphenations drops to 818, with 772 cases involving final vowels, as mentioned earlier. Among these, only two cases involve a final vowel with a diacritic: *fu*zu-ê* and *ba*ba*la-ô*. The list of the remaining 46 follows:

ru-im	*re-ur*ba*ni*zar*	*a*bra-â*mi*co*	*in*flu-ên*cia*	*quar-t.zo*
*flu-i*dez*	*voy-eu*ris*ta*	*ca*far*na-um*	*de*po-i*men*to*	*es*pon*sa-is*
*a*da-il*	*ba-ha*men*se*	*ku-wai*ti*a*no*	*re-u*nir*	*ba*la-us*tra*da*
*voy-eu*ris*mo*	*dar-wi*nis*ta*	*ma*ru-im*	*co-i*bir*	*con*gru-ên*cia*
*a.b-rup*to*	*ma-çô*ni*co*	*su.b-lin*gual*	*flu-ên*cia*	*tran*se-un*te*
*me*ga-watt*	*pat-chu*li*	*te-ô*ni*mo*	*con*flu-ên*cia*	*di*flu-ên*cia*
ca-im	*su.b-li*te*ra*tu*ra*	*a*lu-i*men*to*	*a*nu-ên*cia*	*in*de-is*cen*te*
*su.b-li*mi*nar*	*fa-im*	*mal-*	*su.b-li*nha*	
*dar-wi*nis*mo*	*voy-eu*rís*ti*co*	*thu*si*a*nis*mo*	*a*flu-ên*cia*	
*quar-t.zi*to*	*tai-wa*nês*	*mal-thu*si*a*no*	*in*con*gru-ên*cia*	

[20] A better rule would be to restrain the final hyphenation, leaving the vowel alone, if the preceding syllable has a vowel with diacritic. That would require an exhaustive list of all scenarios.

Table 3: Hyphenation in the dictionaries of words having conflict between morphological and syllabic information.

word	hyphenation	dictionary					
		Michaelis	Priberam	Wikitionary	Aulete	Portal	Dicio
sublinhar	su-bli-nhar		x	x	x		
	sub-li-nhar	x				x	
reiniciar	rei-ni-ci-ar			x		x	
	re-i-ni-ci-ar	x	x		x		x
ciberespaço	ci-be-res-pa-ço		x			x	
	ci-ber-es-pa-ço	x		x	x		x
hiperalgesia	hi-pe-ral-ge-si-a	x	x		x	x	x
	hi-per-al-ge-si-a			x			
autoimagem	au-toi-ma-gem					x	
	au-to-i-ma-gem	x			x		x

From this list, we may create a new rule: `u1ê`, which would also require two exception rules: `1gu2ê` and `1qu2ê`. Incorporating them, we solve the nine cases of missing hyphen *u-ê*, dropping the number of missing hyphenations to 809.

We now move to the last dictionary file: `hyphenations4`. Using the rules we have developed so far, we achieve 0 incorrect and 29 missing hyphenations in this dictionary. Of these, 26 involve final vowels, and the remaining 3 are: *re-in*de*xa*ção*, *pa*ra-i*ba*no* and *ru-ther*ford*)

Considering some additional words that are not on the list, we gather an additional set of rules to be incorporated: `1çô`, and `1xô` (e.g., *ma-çô*ni*co*, *a-xô*nio*, *sa-xô*ni*co*; see rule 23); `a1â`, `a1ã`, `a1ô`, `e1ô`, `o1ã`, and `o1ê` (e.g., *a*bra-â*mi*co*, *a*bra-ão*, *ca*na-ã*, *fa*ra-ô*ni*co*, *le-ô*ni*das*, *na*po*le-ô*ni*co*, *jo-ão*, *la*go-ão*, *bo-ê*mio*; see rule 22); `.ne4o` (e.g., *ne.o*li*be*ral*, *ne.o-na*zis*ta*; see rule 12), `o1im` (e.g., *co-im*bra*; this rule requires the exception rule `co2ima` to avoid unintended errors in words like *coi-ma*; see rule 21), `o2i1na` (e.g., *bo.i*na*, *co.i*na*; created as an exception for `o1in`, see rule 21), and `pro1i1b` (e.g., *pro-i*bir*, see rule 20). After incorporating those rules, the number of incorrect and missing hyphenations, respectively, will be: 0, 538 (`hyphenations6`); 21, 818 (`hyphenations5`); and 0, 33 (`hyphenations4`).

11 Limitations

In general, although the erroneous words share some common characteristics that might allow for a reduction in hyphenation errors to some extent, there is a limitation inherent in the way TEX's rules are conceived. Below, we present the systematics found in these data that could be encompassed by a different rule structure:

Morphological determination Prefixes such as *re-*, *sub-*, *ciber-*, *hiper-*, and *auto-*, among others, require separating the prefix from its stem, which can lead to phonological issues. For example, words like *reiniciar*, *sublinhar*, *ciberespaço*, *hiperalgesia*, and *autoimagem* contain prefixes and could be hyphenated as *re-i-ni-ci-ar*, *sub-li-nhar*, *ci-ber-es-pa-ço*, *hi-per-al-ge-si-a*, and *au-to-i-ma-gem*, respectively, to respect their morphological formation. However, considering that Portuguese hyphenates its words based on syllabic phonological correlates, words that require morphological information, such as these, might not have their hyphenation performed correctly. These examples are presented in Table 3.

Foreignness There is a group of words in the corpus that are terminologies or words incorporated into the Portuguese language without full phonological adaptation, such as *darwinismo*, *quilowatt*, and *esfiha*. The lack of adaptation makes the phonological pattern very specific to the word, making it impossible to incorporate their cases TEX's rules alongside the other rules. The solution is to add them to the exception word list.

Word-initial consonant clusters Portuguese has few cases of consonant clusters at the beginning of a word. They are, in general, etymological remnants and are currently unproductive in the language, since there are no neologisms with this pattern. Encounters like *ps-* and *pn-* are more frequent, as they are present in words like *psicologia* and *pneu*, which have moderate frequency in the language. These can be predicted by specific rules that would cover 49 and 13 words in the corpus, respectively. However, there are consonant clusters that are found in very specific and low-frequency words. Although possible, it is not worth adding very specific rules for clusters found in words like *dzeta*, *gnu*, *cnidário*, *ftálico*, and *gnaisse* – which amount to only five words.

Abbreviations, Acronyms, or Initialisms Whether for efficiency, convenience, clarity, or specialized jargon, it is common to use shortened versions of words or phrases. *Abbreviation* is a method employed to

achieve this shortening. In our Portuguese corpus, we find examples such as *etc.*[21], *Dr.*[22], *Exmo.*[23], *cap.*[24], *Univ.*[25], *ed.*[26], *s.n.*[27]. Another shortened form is an *initialism*, which consists of using the initial letters of words to create a shortened version. However, initialisms may not always conform to the hyphenation rules described in this work, as they do not necessarily follow the orthographic or phonotactic standards of the language. Some abbreviations found in the corpus include *SESC*[28], *INSS*[29], *PCdoB*[30], *PM*[31], and *UFRJ*[32]. *Acronyms* are a specific type of shortening where the first letters (or groups of letters) of each word are combined to form a new pronounceable word. In the corpus, we encounter examples like *Anatel*[33], *Ovni*[34], *Sida*[35] (in Portugal) and *Mercosul*[36]. These various shortened forms play an important role in written language, providing concise ways to represent longer words or phrases. It is important to note that abbreviations, acronyms, and initialisms are generally treated as single units and are not hyphenated.

12 Creating a new set of rules using Patgen

In this section, we describe the process of creating a new set of hyphenation rules using *patgen*. This involves defining specific parameters, choosing a reference dictionary with correct hyphenations, and using a translation file tailored for *patgen*. In the previous Section 3 we have detailed how *patgen* works.

A partial of the content of the translation file is shown in Listing 4.

Listing 4: Portuguese translation file.

```
 1  1 1
 2 %% This file portuguese.tra defines the letters used for generating
 3 %% Portuguese hyphenation patterns with patgen.
 4  a A
 5  á Á
 6  à À
 7  â Ã
 8  ã Ã
 9  b B
10  c C

   ...

39  x X
40  y Y
41  z Z
```

The parameters mentioned in Section 3 to run *patgen* are provided interactively during the execution of *patgen*. We might automatize this process using a script. The Listing 5 presents the code block responsible for passing these parameters to *patgen*.

Listing 5: Portuguese translation file.

```
{
    printf "${hyph_start}␣${hyph_finish}\n"
    for ((i=hyph_start; i<=hyph_finish; i++)); do
        printf "${pat_start}␣${pat_finish}\n"
        printf "${weights}\n"
    done
    printf "y"
} | patgen ../data/portuguese.dic "$filename" "output_${identifier}" ../data/portuguese.
    tra | tail -n 2;
```

[21] Latin expression *et cetera*, meaning 'and other similar things'.
[22] *doutor* (doctor, person with PhD title, but popularly used to designate an erudite individual)
[23] *Excelentíssimo* (honourable)
[24] *capítulo* (chapter)
[25] Universidade (university)
[26] *edição* (edition)
[27] sine nomine, Latin expression meaning 'without a name', mostly used in the context of publishing.
[28] Serviço Social do Comércio
[29] *Instituto Nacional do Seguro Social* (National Institute of Social Security)
[30] *Partido Comunista do Brasil* (Communist Party of Brazil)
[31] *Polícia Militar* (military police)
[32] Universidade Federal do Rio de Janeiro
[33] *Agência Nacional de Telecomunicações* (National Telecommunications Agency)
[34] *Objeto voador não identificado* (unidentified flying object - UFO)
[35] *Síndrome da Imunodeficiência Adquirida* (acquired immunodeficiency syndrome - AIDS)
[36] *Mercado Comum do Sul* (Southern Common Market).

As an example of execution, when choosing `hyph_start`=1, `hyph_finish`=9, `pat_start`=2, `pat_finish`=5, `good weight`=1, `bad weight`=1, and `threshold`=1, we achieve 99.97% (86799) correct hyphenations, 0.01% (12) incorrect hyphenations, and 0.03% (25) missing hyphenations. It produced a set of 799 rules, where just 123 are present in our handcrafted set of rules.

We have also used the approach used by P. Sojka and O. Sojka (2019) in creating the script `make-full-pattern`. In this approach, the parameters are loaded from a file and might be distinct for each level. We start from an empty set and create progressive higher order rules, always using the rules from the previous step as the starting point at the next step. Using this approach with the 8 level parameters they used to create hyphenation rules to German, we achieved 99.98% (86820) correct hyphenations, 0.01% (9) incorrect hyphenations, and 0.00% (4) missed hyphenations. The result was a set of 593 rules, with only 69 in common with those handcrafted rules proposed earlier.

The results from the set of rules crafted by *patgen* seems outstanding, but we must be skeptical of its generalization power. To inquire on this matter, we decided to retrieve the list of words where just three hyphenations were found in the dictionaries. We have also dealt with apparent proparoxytones, to be consistent in our approach. In this new set, the number of correct, incorrect and missing hyphenations for each set of rules is given Table 4.

Table 4: Compared results between two use cases of *patgen* and the handcrafted rule set when hyphenating the set of words from the dictionary built from those words with 3 agreements on the online dictionaries.

rule set	correct	incorrect	missing
fixed parameters	7465	150	183
make-full-pattern	647	47	7089
handcrafted	7335	42	385

It is important to hightlight that, among those 385 missing hyphenations in the handcrafted rules, 278 cases correspond to the last hyphen, leaving a single vowel as a final syllable. Comparing to the other set of rules, we have 7 (in 183, for fixed parameters case), and 257 (in 7089, for the `make-full-pattern` script).

Establishing this comparison leave us no doubt that our handcrafted rules achieved the best performance, also retaining a small number of useful rules, what leads to good generalization.

13 Hyphenation rule patches

We systematize the set of new rules in this section, and provide a few examples for each instance. These rules are intended as a complement for the default TEX hyphenation rules created by Rezende and Almeida 2015.

1 rule: `.g2no`, `.g2nó`, `.g2nô` – *gnomo*, *gnóstico*, *gnômon*

2 rule: `t2c` – *tchau*, *tcheco*

3 rule: `1p2neu` – *pneumonia*, *pneumotórax*, *pneumático*, *hidropneumático*

4 rule: `.t2m` – *tmese*

5 rule: `.p2t` – *ptose*, *pterossauro*

6 rule: `.m2n` – *mnemônico*

7 rule: `c2za` – *czar*

8 rule: `.s2` – *stalinismo*

9 rule: `.t2` – tsunami, tzarista

10 rule: `.p2si`, `.p2sí` – *psicologia*, *psíquico*

11 rule: `su2b3r`, `su2b3l` – *sublunar*, *subrotina*
exception: `.su3b4li` – *sublinhar*, *sublimar*

12 rule: `.ne4o` – *neoliberal*, *neonazista*

13 rule: `1p2seu1d` – *pseudônimo*

14 rule: `1qu` – *enquanto*, *inquieto*, *farquhar*, *qubit*

15 rule: `a1ir.`, `u1ir.` – *sair*, *extrair*, *diminuir*, *incluir*

16 rule: `a1ind`, `a1i1nh` – *ainda*, *rainha*

17 rule: `e1imp` – *reimpresso*, *teleimpressor*

18 rule: `e1inc`, `e1inf`, `e1ing`, `e1ins`, `e1int`, `e1inv` – *reincidência*, *reinfecção*, *reingressa*, *reinserção*, *reintegração*, *reinventar*

19 rule: `u1iz.`, `a1iz.` – *juiz*, *raiz*

20 rule: `pro1i1b` – proibição

21 rule: `tu1i`, `bu1i`, `nu1i`, `o1im`, `o1in`, `u1in`, `su1i`, `í1e`, `ju1i`, `fu1i`, `du1i`, `do1im`, `au1i`, `u1i1ç` – *intuitivo*, *contribuidor*, *ingenuidade*, *coimbra*, *coincide*, *ruindade*, *suicida*, *píer*, *juizado*, *fuinha*, *assiduidade*, *amendoim*, *cacauicultor*, *constituição*
exception: `tu2id`, `tu2it`, `co2ima`, `o2i1na` – *gratuidade*, *intuito*, *coima*, *boina*

22 rule: `a1â`, `a1ã`, `a1é`, `a1í`, `a1ó`, `a1ô`, `a1ú`, `e1á`, `e1â`, `e1ã`, `e1é`, `e1ê`, `e1í`, `e1ó`, `e1ô`, `e1ú`, `é1o`, `i1á`, `i1ã`, `i1é`, `i1í`, `i1ó`, `i1u`, `i1ú`, `í1a`, `í1o`, `o1á`, `o1ã`, `o1é`, `o1ê`, `o1í`, `o1ó`, `u1á`, `u1ã`, `u1â`, `u1é`, `u1ê`, `u1í`, `ú1o` – *abraâmico*, *abraão*, *aéreo*, *país*, *caótico*, *faraônico*, *saúde*, *balneário*, *oceânico*, *campeã*, *feérico*, *veêm*, *veículo*, *teórico*, *napoleônico*, *conteúdo*, *néon*, *diário*, *região*, *soviético*, *iídiche*, *periódico*, *feiura*, *viúva*, *maníaco*, *íon*, *razoável*, *joão*, *poético*, *boêmia*, *heroísmo*, *alcoólico*, *usuário*, *itapuã*, *lituânia*, *suécia*, *cauê*, *suíça*, *flúor*
exception: `1gu2á`, `1gu2ã`, `1gu2é`, `1gu2ê`, `1gu2í`, `1qu2á`, `1qu2ã`, `1qu2â`, `1qu2é`, `1qu2ê`, `1qu2í`, – *jaraguá*, *saguão*, *alguém*, *português*, *linguística*, *aquático*, *camaquã*, *equânime*, *inquérito*, *sequência*, *química*

23 rule: `1bô`, `1cô`, `1çô`, `1dô`, `1fô`, `1gô`, `1lô`, `1pô`, `1mô`, `1nô`, `1rô`, `1sô`, `1tô`, `1vô`, `1xô`, `1zô` – *robô*, *recôncavo*, *maçônico*, *judô*, *telefônica*, *xangô*, *camelô*, *capô*, *sumô*, *econômico*, *subsônico*, *tarô*, *chatô*, *vovô*, *saxônia*, *amazônia*

24 rule: `4a.`, `4e.`, `4o.` – *secretária*, *planície*, *paratormônio*

The 120 rules were grouped above in a list of 24 types of rules. They may be further organized into five large groups. The first, which comprises rules 1 to 9, includes consonant clusters such as *czar*, *ptose* and *gnomo*. The second group, comprising rules 10 to 13, delimits the morphological boundary between prefixes and radicals. As noted, although phonological issues guide the separation of numerous words in Portuguese, there are also those that are guided by morphology. This is the case of words that have the prefixes *sub-* and *re-*, such as *sublunar* and *reinserção*. The third group, comprising rules 14 to 22, seeks to understand a set of words that have vowel combinations that do not follow the general rules. This is because the Portuguese language has vowel encounters with the second vowel graphically marked that can be separated, forming hiatuses, such as *caótico*, *balneário* and *razoável*, while there are also words with a similar structure that constitute a diphthong, such as *português*, *alguém* and *linguística*. It is remarkable, of course, that the latter are formed by the digraphs *qu-* and *gu-*, while the former by vowels other than *i* and *u*. The fourth group, in turn, which comprises rules 22 and 23[37], which are counterparts of rules that were already in the default rules, but did not contemplate the cases with certain accents. They were then added to encompass words such as *camelô*, *recôncavo*, *amazônia*, and *maçônico*. The fifth, and last group, has just a single instance, rule 24, which represents our choice in how to deal the apparent proparoxytones, avoiding a final hyphenation with the vowels *a*, *e*, or *o*.

The set made of 120 rules is presented below:

```
.p2si       1tô       i1í       e1imp      1fô       u1ê
.p2sí       1rô       i1ó       e1inc      1mô       1gu2ê
.g2no       1pô       í1o       e1inf      1nô       1qu2ê
.g2nó       a1é       i1u       e1ing      1sô       1çô
.g2nô       a1í       i1ú       e1ins      1zô       u1é
t2c         a1ó       o1á       e1int      tu1i      1gu2é
1p2neu      a1ú       o1é       e1inv      tu2it     1qu2é
.t2m        e1á       o1í       u1iz.      tu2id     1xô
.p2t        e1â       o1ó       a1iz.      bu1i      a1â
su2b3r      e1ã       u1á       4a.        nu1i      a1ã
su2b3l      e1é       u1ã       4e.        o1in      a1ô
.su3b4li    e1ê       u1â       4o.        u1in      e1ô
a1ir.       e1í       u1í       1gu2á      su1i      .ne4o
u1ir.       e1ó       ú1o       1gu2ã      í1e       o1ã
1qu         é1o       1qu2á     1qu2ã      ju1i      o1ê
1vô         e1ú       1qu2â     .m2n       fu1i      o1im
1lô         i1á       1qu2í     c2za       du1i      o2i1na
1cô         i1ã       1gu2í     .s2        do1im     pro1i1b
1gô         í1a       a1ind     1p2seu1d   au1i      co2ima
1bô         i1é       a1i1nh    1dô        u1i1ç     .t2
```

[37]Note that rule 22 is in the intersection between the third and fourth group of rules.

Table 5 presents the hyphenation errors in a set of 876 words. We have omitted those words where both rules let to a correct hyphenation and those that have incorrect hyphenation in a vowel in the last position. Once again, we are using the *patgen* notation to mark correct, incorrect, and missing hyphenation points: `*`, `.`, and `-`, respectively. Incorrect or missing hyphenation points are considered errors. In these cases, we place mark ✗. Correct hyphenations are marked with ✓. The default rules achieved 17 correct hyphenations in this set, against 811 of the patched rules.

14 Concluding Remarks and Future Directions

In this study, we explored the default TeX hyphenation rules and proposed additional rules to improve hyphenation performance in Portuguese. By addressing the limitations of the old rules and existing approaches within the TeX typesetting system, and by incorporating morphological and phonological considerations, we significantly reduced the number of hyphenation errors.

However, the patched set of rules is still not enough to achieve perfect accuracy, and it is not even desired since our dataset is noisy and there are many dubious hyphenation cases. We have opted to create a concise set of rules that could better generalize and, therefore, align more closely with the underlying hyphenation rules in the language.

We also tested rules created by patgen, which generated an extensive set of rules that were unable to generalize effectively. In contrast, our handcrafted rules performed superiorly in a validation set.

In conclusion, while our enhanced rule set demonstrates significant improvements in hyphenation accuracy, there could remains room for further refinement.

Wide character support and universal syllabic segmentation, as considered by O. Sojka, P. Sojka, and Máca 2023, are additional points for consideration. Incorporating regular expressions, creating classes of characters, and supporting word stress positions might further enhance the efficiency and generality of the hyphenation system.

References

Araújo, Antonio Martins de and Toru Maruyama (2015). "A Hifenização em Português". In: *Idioma* 28, pp. 90–107. URL: `http://www.institutodeletras.uerj.br/idioma/numeros/28/Idioma28_a08.pdf`.

Bergström, Magnus and Neves Reis (2011). *Prontuário ortográfico e guia da língua portuguesa*. Leya.

Cagliari, Luiz Carlos (2015). "Aspectos teóricos da ortografia". In: *Ortografia da língua portuguesa: história, discurso, representações*. Ed. by Maurício Silva. São Paulo: Contexto, pp. 17–52.

Cegalla, Domingos Paschoal (2020). *Novíssima Gramática Da Línngua Portuguesa*. São Paulo: Companhia Editora Nacional.

Coulmas, Florian (2003). *Writing Systems: An Introduction to Their Linguistic Analysis*. Cambridge Textbooks in Linguistics. Cambridge University Press. ISBN: 9780521787376.

Cunha, Celso and Lindley Cintra (2016). *Nova gramática do português contemporâneo*. LEXIKON Editora Digital ltda.

Gundlach, Patrick (2021). *A port of TeX's hyphenation algorithm to Go*. URL: `https://github.com/speedata/hyphenation`.

Haralambous, Yannis (2021). "A Revisited Small Tutorial on Patgen, 28 Years After". In: *electronic form, available from CTAN as info/patgen2. tutorial*.

Honorof, Douglas and Laurie Feldman (Jan. 2006). "The Chinese character in psycholinguistic research: Form, structure, and the reader". In: *The Handbook of East Asian Psycholinguistics*. Ed. by Ping Li et al. Vol. 1. Cambridge University Press. Chap. 17, pp. 195–208. DOI: `10.2277/0521833337`.

Hunspell's Team (Apr. 13, 2023a). *Hunspell*. URL: `http://hunspell.github.io/`.

— (Apr. 13, 2023b). *Hunspell Hyphen*. URL: `https://github.com/hunspell/hyphen`.

Knuth, Donald Ervin (May 13, 1977). *Preliminary preliminary description of TEX*. unpublished.

Kristensen, T. and D. Langmyhr (2001). "Two regimes of computer hyphenation - a comparison". In: *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*. IEEE. DOI: `10.1109/ijcnn.2001.939592`. URL: `https://doi.org/10.1109/ijcnn.2001.939592`.

Levien, Raph (1998). *Brief explanation of the hyphenation algorithm herein*. Hunspell. URL: `https://github.com/hunspell/hyphen/blob/master/README.hyphen`.

Liang, Frank and Peter Breitenlohner (1991). *PATtern GENeration program for the TEX82 hyphenator*. Tech. rep. 2. Electronic documentation of PATGEN.

Liang, Franklin Mark (1983). "Word Hy-phen-a-tion by Com-put-er". PhD thesis. Stanford University.

Libossek, Marion and Florian Schiel (2000). "Syllable-based text-to-phoneme conversion for German." In: *INTERSPEECH*. Citeseer, pp. 283–286.

Lin, Li-chin (2011). “Fundamental generalizations of English syllabification”. In: *Concentric: Studies in Linguistics* 37.2, pp. 179–208.
Magalhães Gândavo, Pêro de (1574). *Regras que ensinam a maneira de escrever e orthographia da lingoa portuguesa, etc.* Lisboa. URL: https://purl.pt/324.
Mateus, Maria Helena Mira (n.d.). “Questões fonológicas do português”. manuscript.
Németh, László (2006). “Automatic non-standard hyphenation in OpenOffice. org”. In: *TUGboat* 27.1, pp. 32–37.
Palmer, David D. (2010). “Tokenisation and Sentence Segmentation”. In: *Handbook of Natural Language Processing*. Ed. by Nitin Indurkhya and Fred J. Damerau. CRC Press. Chap. 2, pp. 9–30.
TEX pattern authors (Apr. 13, 2023). *TEX hyphenation patterns*. URL: https://www.tug.org/tex-hyphen/.
Rezende, Pedro Jussieu de (1987). “Portuguese hyphenation table for TEX”. In: *TUGboat* 8.2, pp. 102–102.
Rezende, Pedro Jussieu de and José Joao Dias Almeida (2015). *Hyphenation patterns for Portuguese*. http://mirror.ctan.org/language/hyph-utf8/tex/generic/hyph-utf8/patterns/tex/hyph-pt.tex.
Scannell, Kevin (2003). “Hyphenation patterns for minority languages”. In: *TUGboat* 24.2, pp. 236–239.
Sojka, Ondřej, Petr Sojka, and Jakub Máca (2023). “A roadmap for universal syllabic segmentation”. In: *TUGboat* 44.2, pp. 289–296. DOI: 10.47397/tb/44-2/tb137sojka-syllabic.
Sojka, Petr (Aug. 1995a). *Notes on Compound Word Hyphenation in TEX*. Tech. rep. Masaryk University in Brno, Faculty of Informatics.
— (1995b). “Notes on compound word hyphenation in TEX”. In: *TUGboat* 16.3, pp. 290–297.
Sojka, Petr et al. (2005). “Competing Patterns in Language Engineering and Computer Typesetting”. PhD thesis. Masaryk University, Brno.
Sojka, Petr and David Antoš (2003). “Context sensitive pattern based segmentation: A Thai challenge”. In: *Proceedings of EACL 2003 Workshop on Computational Linguistics for South Asian Languages–Expanding Synergies with Europe, Budapest*, pp. 65–72.
Sojka, Petr and Ondřej Sojka (2019). “The unreasonable effectiveness of pattern generation”. In: *TUGboat* 40.2, pp. 187–193.
Trogkanis, Nikolaos and Charles Elkan (2010). “Conditional random fields for word hyphenation”. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 366–374.
Yavas, Mehmet (2020). *Applied English Phonology*. John Wiley & Sons.

A Hyphenation results

Table 5: Comparison of hyphenation results in a sample of words using the default rules and the proposed patched rules. In this sample we select cases where there are incorrect hyphenations from either set of rules and we are discarding those cases with incorrect hyphenation before a final vowel.

word	defult rule result	patched rules result
aarônico	a*a-rô*ni*co ✗	a*a*rô*ni*co ✓
abdômen	ab-dô*men ✗	ab*dô*men ✓
abiótico	a*bi-ó*ti*co ✗	a*bi*ó*ti*co ✓
abluir	a*blu-ir ✗	a*blu*ir ✓
abraâmico	a*bra-â*mi*co ✗	a*bra*â*mi*co ✓
abrupto	a.b-rup*to ✗	a.b-rup*to ✗
absenteísmo	ab*sen*te-ís*mo ✗	ab*sen*te*ís*mo ✓
abstraído	abs*tra-í*do ✗	abs*tra*í*do ✓
abstrair	abs*tra-ir ✗	abs*tra*ir ✓
acetaldeído	a*ce*tal*de-í*do ✗	a*ce*tal*de*í*do ✓
acônito	a-cô*ni*to ✗	a*cô*ni*to ✓
acordeão	a*cor*de-ão ✗	a*cor*de*ão ✓
acuidade	a*cu-i*da*de ✗	a*cu-i*da*de ✗
adail	a*da-il ✗	a*da-il ✗
adriático	a*dri-á*ti*co ✗	a*dri*á*ti*co ✓
aético	a-é*ti*co ✗	a*é*ti*co ✓
afluir	a*flu-ir ✗	a*flu*ir ✓
afrodisíaco	a*fro*di*sí-a*co ✗	a*fro*di*sí*a*co ✓
agogô	a*go-gô ✗	a*go*gô ✓
agrião	a*gri-ão ✗	a*gri*ão ✓
agronômico	a*gro-nô*mi*co ✗	a*gro*nô*mi*co ✓
ainda	a-in*da ✗	a*in*da ✓
ajuizar	a*ju-i*zar ✗	a*ju*i*zar ✓
alaúde	a*la-ú*de ✗	a*la*ú*de ✓

word	defult rule result	patched rules result
alcíone	al*cí-o*ne ✗	al*cí*o*ne ✓
alcoólatra	al*co-ó*la*tra ✗	al*co*ó*la*tra ✓
alcoólico	al*co-ó*li*co ✗	al*co*ó*li*co ✓
aldeão	al*de-ão ✗	al*de*ão ✓
aldeído	al*de-í*do ✗	al*de*í*do ✓
aleúte	a*le-ú*te ✗	a*le*ú*te ✓
aliás	a*li-ás ✗	a*li*ás ✓
almôndega	al-môn*de*ga ✗	al*môn*de*ga ✓
alô	a-lô ✗	a*lô ✓
aloés	a*lo-és ✗	a*lo*és ✓
alquímico	al-quí*mi*co ✗	al*quí*mi*co ✓
altruísmo	al*tru-ís*mo ✗	al*tru*ís*mo ✓
altruísta	al*tru-ís*ta ✗	al*tru*ís*ta ✓
aluimento	a*lu-i*men*to ✗	a*lu-i*men*to ✗
aluir	a*lu-ir ✗	a*lu*ir ✓
aluvião	a*lu*vi-ão ✗	a*lu*vi*ão ✓
alvéolo	al*vé-o*lo ✗	al*vé*o*lo ✓
amazônico	a*ma-zô*ni*co ✗	a*ma*zô*ni*co ✓
amebíase	a*me*bí-a*se ✗	a*me*bí*a*se ✓
amendoim	a*men*do-im ✗	a*men*do*im ✓
aminoácido	a*mi*no-á*ci*do ✗	a*mi*no*á*ci*do ✓
amiúde	a*mi-ú*de ✗	a*mi*ú*de ✓
amoníaco	a*mo*ní-a*co ✗	a*mo*ní*a*co ✓
anafrodisíaco	a*na*fro*di*sí-a*co ✗	a*na*fro*di*sí*a*co ✓
anatômico	a*na-tô*mi*co ✗	a*na*tô*mi*co ✓
ancião	an*ci-ão ✗	an*ci*ão ✓
ancilostomíase	an*ci*los*to*mí-a*se ✗	an*ci*los*to*mí*a*se ✓
anemômetro	a*ne-mô*me*tro ✗	a*ne*mô*me*tro ✓
anfitrião	an*fi*tri-ão ✗	an*fi*tri*ão ✓
anônimo	a-nô*ni*mo ✗	a*nô*ni*mo ✓
antagônico	an*ta-gô*ni*co ✗	an*ta*gô*ni*co ✓
antiácido	an*ti-á*ci*do ✗	an*ti*á*ci*do ✓
antialcoólico	an*ti*al*co-ó*li*co ✗	an*ti*al*co*ó*li*co ✓
antibiótico	an*ti*bi-ó*ti*co ✗	an*ti*bi*ó*ti*co ✓
antipatriótico	an*ti*pa*tri-ó*ti*co ✗	an*ti*pa*tri*ó*ti*co ✓
antiquíssimo	an*ti-quís*si*mo ✗	an*ti*quís*si*mo ✓
antônimo	an-tô*ni*mo ✗	an*tô*ni*mo ✓
antropônimo	an*tro-pô*ni*mo ✗	an*tro*pô*ni*mo ✓
anuidade	a*nu-i*da*de ✗	a*nu*i*da*de ✓
anuir	a*nu-ir ✗	a*nu*ir ✓
aórtico	a-ór*ti*co ✗	a*ór*ti*co ✓
apreciável	a*pre*ci-á*vel ✗	a*pre*ci*á*vel ✓
aquático	a-quá*ti*co ✗	a*quá*ti*co ✓
aquém	a-quém ✗	a*quém ✓
aquícola	a-quí*co*la ✗	a*quí*co*la ✓
aquífero	a-quí*fe*ro ✗	a*quí*fe*ro ✓
arábias	a*rá*bi.as ✗	a*rá*bi.as ✗
arcaísmo	ar*ca-ís*mo ✗	ar*ca*ís*mo ✓
aríete	a*rí-e*te ✗	a*rí*e*te ✓
arqueólogo	ar*que-ó*lo*go ✗	ar*que*ó*lo*go ✓
arquétipo	ar-qué*ti*po ✗	ar*qué*ti*po ✓
arquidiácono	ar*qui*di-á*co*no ✗	ar*qui*di*á*co*no ✓
arquitetônica	ar*qui*te-tô*ni*ca ✗	ar*qui*te*tô*ni*ca ✓
arquitetônico	ar*qui*te-tô*ni*co ✗	ar*qui*te*tô*ni*co ✓
arrais	ar*ra-is ✗	ar*ra-is ✗
arruinar	ar*ru-i*nar ✗	ar*ru*i*nar ✓
ascaridíase	as*ca*ri*dí-a*se ✗	as*ca*ri*dí*a*se ✓
asiática	a*si-á*ti*ca ✗	a*si*á*ti*ca ✓
asiático	a*si-á*ti*co ✗	a*si*á*ti*co ✓
assiduidade	as*si*du-i*da*de ✗	as*si*du*i*da*de ✓
associável	as*so*ci-á*vel ✗	as*so*ci*á*vel ✓
astronômico	as*tro-nô*mi*co ✗	as*tro*nô*mi*co ✓
ataúde	a*ta-ú*de ✗	a*ta*ú*de ✓
ateísmo	a*te-ís*mo ✗	a*te*ís*mo ✓
ateísta	a*te-ís*ta ✗	a*te*ís*ta ✓

word	defult rule result	patched rules result
atômico	a-tô*mi*co ✗	a*tô*mi*co ✓
atraído	a*tra-í*do ✗	a*tra*í*do ✓
atrair	a*tra-ir ✗	a*tra*ir ✓
atribuição	a*tri*bu-i*ção ✗	a*tri*bu*i*ção ✓
atribuir	a*tri*bu-ir ✗	a*tri*bu*ir ✓
atribuível	a*tri*bu-í*vel ✗	a*tri*bu*í*vel ✓
auréola	au*ré-o*la ✗	au*ré*o*la ✓
austríaco	aus*trí-a*co ✗	aus*trí*a*co ✓
autônimo	au-tô*ni*mo ✗	au*tô*ni*mo ✓
autônomo	au-tô*no*mo ✗	au*tô*no*mo ✓
avião	a*vi-ão ✗	a*vi*ão ✓
avô	a-vô ✗	a*vô ✓
babuíno	ba*bu-í*no ✗	ba*bu*í*no ✓
bacteriófago	bac*te*ri-ó*fa*go ✗	bac*te*ri*ó*fa*go ✓
bafômetro	ba-fô*me*tro ✗	ba*fô*me*tro ✓
bahamense	ba-ha*men*se ✗	ba-ha*men*se ✗
baião	bai-ão ✗	bai*ão ✓
bainha	ba-i*nha ✗	ba*i*nha ✓
baiuca	bai-u*ca ✗	bai*u*ca ✓
balaustrada	ba*la-us*tra*da ✗	ba*la-us*tra*da ✗
balaústre	ba*la-ús*tre ✗	ba*la*ús*tre ✓
bangalô	ban*ga-lô ✗	ban*ga*lô ✓
bastião	bas*ti-ão ✗	bas*ti*ão ✓
batuíra	ba*tu-í*ra ✗	ba*tu*í*ra ✓
benjoim	ben*jo-im ✗	ben*jo*im ✓
bibelô	bi*be-lô ✗	bi*be*lô ✓
bibliófilo	bi*bli-ó*fi*lo ✗	bi*bli*ó*fi*lo ✓
bicampeão	bi*cam*pe-ão ✗	bi*cam*pe*ão ✓
bioética	bi*o-é*ti*ca ✗	bi*o*é*ti*ca ✓
biófilo	bi-ó*fi*lo ✗	bi*ó*fi*lo ✓
biógrafo	bi-ó*gra*fo ✗	bi*ó*gra*fo ✓
biólogo	bi-ó*lo*go ✗	bi*ó*lo*go ✓
bioquímica	bi*o-quí*mi*ca ✗	bi*o*quí*mi*ca ✓
biótico	bi-ó*ti*co ✗	bi*ó*ti*co ✓
biótipo	bi-ó*ti*po ✗	bi*ó*ti*po ✓
biquíni	bi-quí*ni ✗	bi*quí*ni ✓
birô	bi-rô ✗	bi*rô ✓
bisavô	bi*sa-vô ✗	bi*sa*vô ✓
biscainho	bis*ca-i*nho ✗	bis*ca*i*nho ✓
bocaiuva	bo*cai-u*va ✗	bo*cai*u*va ✓
boião	boi-ão ✗	boi*ão ✓
boiuna	boi-u*na ✗	boi*u*na ✓
bongô	bon-gô ✗	bon*gô ✓
bordô	bor-dô ✗	bor*dô ✓
brechtiano	bre.ch*ti*a*no ✗	bre.ch*ti*a*no ✗
bronquíolo	bron-quí-o*lo ✗	bron*quí*o*lo ✓
buir	bu-ir ✗	bu*ir ✓
buquê	bu-quê ✗	bu*quê ✓
cafarnaum	ca*far*na-um ✗	ca*far*na-um ✗
cafeína	ca*fe-í*na ✗	ca*fe*í*na ✓
caíco	ca-í*co ✗	ca*í*co ✓
caída	ca-í*da ✗	ca*í*da ✓
caído	ca-í*do ✗	ca*í*do ✓
caim	ca-im ✗	ca-im ✗
caíque	ca-í*que ✗	ca*í*que ✓
cair	ca-ir ✗	ca*ir ✓
cajuína	ca*ju-í*na ✗	ca*ju*í*na ✓
camaleão	ca*ma*le-ão ✗	ca*ma*le*ão ✓
camelô	ca*me-lô ✗	ca*me*lô ✓
camião	ca*mi-ão ✗	ca*mi*ão ✓
campainha	cam*pa-i*nha ✗	cam*pa*i*nha ✓
campeão	cam*pe-ão ✗	cam*pe*ão ✓
candidíase	can*di*dí-a*se ✗	can*di*dí*a*se ✓
canoísta	ca*no-ís*ta ✗	ca*no*ís*ta ✓
caótico	ca-ó*ti*co ✗	ca*ó*ti*co ✓

word	defult rule result	patched rules result
capô	ca-pô ✗	ca*pô ✓
caraíba	ca*ra-í*ba ✗	ca*ra*í*ba ✓
carbônico	car-bô*ni*co ✗	car*bô*ni*co ✓
cardíaco	car*dí-a*co ✗	car*dí*a*co ✓
cariótipo	ca*ri-ó*ti*po ✗	ca*ri*ó*ti*po ✓
carnaúba	car*na-ú*ba ✗	car*na*ú*ba ✓
caseína	ca*se-í*na ✗	ca*se*í*na ✓
casuísmo	ca*su-ís*mo ✗	ca*su*ís*mo ✓
casuísta	ca*su-ís*ta ✗	ca*su*ís*ta ✓
casuística	ca*su-ís*ti*ca ✗	ca*su*ís*ti*ca ✓
casuístico	ca*su-ís*ti*co ✗	ca*su*ís*ti*co ✓
catequético	ca*te-qué*ti*co ✗	ca*te*qué*ti*co ✓
catião	ca*ti-ão ✗	ca*ti*ão ✓
cauim	cau-im ✗	cau*im ✓
caúna	ca-ú*na ✗	ca*ú*na ✓
centurião	cen*tu*ri-ão ✗	cen*tu*ri*ão ✓
chatô	cha-tô ✗	cha*tô ✓
ciático	ci-á*ti*co ✗	ci*á*ti*co ✓
cirurgião	ci*rur*gi-ão ✗	ci*rur*gi*ão ✓
ciúme	ci-ú*me ✗	ci*ú*me ✓
coágulo	co-á*gu*lo ✗	co*á*gu*lo ✓
cocaína	co*ca-í*na ✗	co*ca*í*na ✓
cocô	co-cô ✗	co*cô ✓
codeína	co*de-í*na ✗	co*de*í*na ✓
coibir	co-i*bir ✗	co-i*bir ✗
coincidente	co-in*ci*den*te ✗	co*in*ci*den*te ✓
coincidir	co-in*ci*dir ✗	co*in*ci*dir ✓
coleóptero	co*le-óp*te*ro ✗	co*le*óp*te*ro ✓
comediógrafo	co*me*di-ó*gra*fo ✗	co*me*di*ó*gra*fo ✓
conciliábulo	con*ci*li-á*bu*lo ✗	con*ci*li*á*bu*lo ✓
conciliável	con*ci*li-á*vel ✗	con*ci*li*á*vel ✓
concluído	con*clu-í*do ✗	con*clu*í*do ✓
concluinte	con*clu-in*te ✗	con*clu*in*te ✓
concluir	con*clu-ir ✗	con*clu*ir ✓
condoído	con*do-í*do ✗	con*do*í*do ✓
confiável	con*fi-á*vel ✗	con*fi*á*vel ✓
confluir	con*flu-ir ✗	con*flu*ir ✓
constituição	cons*ti*tu-i*ção ✗	cons*ti*tu*i*ção ✓
constituído	cons*ti*tu-í*do ✗	cons*ti*tu*í*do ✓
constituinte	cons*ti*tu-in*te ✗	cons*ti*tu*in*te ✓
constituir	cons*ti*tu-ir ✗	cons*ti*tu*ir ✓
construído	cons*tru-í*do ✗	cons*tru*í*do ✓
construir	cons*tru-ir ✗	cons*tru*ir ✓
conteúdo	con*te-ú*do ✗	con*te*ú*do ✓
contraído	con*tra-í*do ✗	con*tra*í*do ✓
contrair	con*tra-ir ✗	con*tra*ir ✓
contribuição	con*tri*bu-i*ção ✗	con*tri*bu*i*ção ✓
contribuidor	con*tri*bu-i*dor ✗	con*tri*bu*i*dor ✓
contribuinte	con*tri*bu-in*te ✗	con*tri*bu*in*te ✓
contribuir	con*tri*bu-ir ✗	con*tri*bu*ir ✓
copaíba	co*pa-í*ba ✗	co*pa*í*ba ✓
coreógrafo	co*re-ó*gra*fo ✗	co*re*ó*gra*fo ✓
coroinha	co*ro-i*nha ✗	co*ro*i*nha ✓
corroído	cor*ro-í*do ✗	cor*ro*í*do ✓
criciúma	cri*ci-ú*ma ✗	cri*ci*ú*ma ✓
criptônimo	crip-tô*ni*mo ✗	crip*tô*ni*mo ✓
cronômetro	cro-nô*me*tro ✗	cro*nô*me*tro ✓
cuíca	cu-í*ca ✗	cu*í*ca ✓
curião	cu*ri-ão ✗	cu*ri*ão ✓
czarina	c.za*ri*na ✗	cza*ri*na ✓
czarismo	c.za*ris*mo ✗	cza*ris*mo ✓
czarista	c.za*ris*ta ✗	cza*ris*ta ✓
dadaísmo	da*da-ís*mo ✗	da*da*ís*mo ✓
dadaísta	da*da-ís*ta ✗	da*da*ís*ta ✓
daguerreótipo	da*guer*re-ó*ti*po ✗	da*guer*re*ó*ti*po ✓

word	defult rule result	patched rules result
daltônico	dal-tô*ni*co ✗	dal*tô*ni*co ✓
darwinismo	dar-wi*nis*mo ✗	dar-wi*nis*mo ✗
darwinista	dar-wi*nis*ta ✗	dar-wi*nis*ta ✗
deão	de-ão ✗	de*ão ✓
decacampeão	de*ca*cam*pe-ão ✗	de*ca*cam*pe*ão ✓
decaída	de*ca-í*da ✗	de*ca*í*da ✓
decaído	de*ca-í*do ✗	de*ca*í*do ✓
decair	de*ca-ir ✗	de*ca*ir ✓
decurião	de*cu*ri-ão ✗	de*cu*ri*ão ✓
deísmo	de-ís*mo ✗	de*ís*mo ✓
deísta	de-ís*ta ✗	de*ís*ta ✓
demiurgo	de*mi-ur*go ✗	de*mi*ur*go ✓
demoníaco	de*mo*ní-a*co ✗	de*mo*ní*a*co ✓
depoimento	de*po-i*men*to ✗	de*po*i*men*to ✓
descaída	des*ca-í*da ✗	des*ca*í*da ✓
descair	des*ca-ir ✗	des*ca*ir ✓
desconstruir	des*cons*tru-ir ✗	des*cons*tru*ir ✓
descontraído	des*con*tra-í*do ✗	des*con*tra*í*do ✓
descontrair	des*con*tra-ir ✗	des*con*tra*ir ✓
desembainhar	de*sem*ba-i*nhar ✗	de*sem*ba*i*nhar ✓
desobstruir	de*sobs*tru-ir ✗	de*sobs*tru*ir ✓
despoluição	des*po*lu-i*ção ✗	des*po*lu*i*ção ✓
despoluir	des*po*lu-ir ✗	des*po*lu*ir ✓
destituição	des*ti*tu-i*ção ✗	des*ti*tu*i*ção ✓
destituído	des*ti*tu-í*do ✗	des*ti*tu*í*do ✓
destituir	des*ti*tu-ir ✗	des*ti*tu*ir ✓
destruição	des*tru-i*ção ✗	des*tru*i*ção ✓
destruído	des*tru-í*do ✗	des*tru*í*do ✓
destruir	des*tru-ir ✗	des*tru*ir ✓
diácono	di-á*co*no ✗	di*á*co*no ✓
díade	dí-a*de ✗	dí*a*de ✓
diáfano	di-á*fa*no ✗	di*á*fa*no ✓
diálise	di-á*li*se ✗	di*á*li*se ✓
diálogo	di-á*lo*go ✗	di*á*lo*go ✓
diáspora	di-ás*po*ra ✗	di*ás*po*ra ✓
diástole	di-ás*to*le ✗	di*ás*to*le ✓
dicroísmo	di*cro-ís*mo ✗	di*cro*ís*mo ✓
diérese	di-é*re*se ✗	di*é*re*se ✓
diferenciável	di*fe*ren*ci-á*vel ✗	di*fe*ren*ci*á*vel ✓
diluído	di*lu-í*do ✗	di*lu*í*do ✓
diluir	di*lu-ir ✗	di*lu*ir ✓
diminuição	di*mi*nu-i*ção ✗	di*mi*nu*i*ção ✓
diminuído	di*mi*nu-í*do ✗	di*mi*nu*í*do ✓
diminuir	di*mi*nu-ir ✗	di*mi*nu*ir ✓
dinamarquês	di*na*mar-quês ✗	di*na*mar*quês ✓
dionisíaco	di*o*ni*sí-a*co ✗	di*o*ni*sí*a*co ✓
dióxido	di-ó*xi*do ✗	di*ó*xi*do ✓
distraído	dis*tra-í*do ✗	dis*tra*í*do ✓
distrair	dis*tra-ir ✗	dis*tra*ir ✓
distribuição	dis*tri*bu-i*ção ✗	dis*tri*bu*i*ção ✓
distribuído	dis*tri*bu-í*do ✗	dis*tri*bu*í*do ✓
distribuidora	dis*tri*bu-i*do*ra ✗	dis*tri*bu*i*do*ra ✓
distribuidor	dis*tri*bu-i*dor ✗	dis*tri*bu*i*dor ✓
distribuir	dis*tri*bu-ir ✗	dis*tri*bu*ir ✓
diurno	di-ur*no ✗	di*ur*no ✓
dodecafônico	do*de*ca-fô*ni*co ✗	do*de*ca*fô*ni*co ✓
doído	do-í*do ✗	do*í*do ✓
dríade	drí-a*de ✗	drí*a*de ✓
dzeta	d.ze*ta ✗	d.ze*ta ✗
eclesiástico	e*cle*si-ás*ti*co ✗	e*cle*si*ás*ti*co ✓
econômico	e*co-nô*mi*co ✗	e*co*nô*mi*co ✓
ecônomo	e-cô*no*mo ✗	e*cô*no*mo ✓
egoísmo	e*go-ís*mo ✗	e*go*ís*mo ✓
egoísta	e*go-ís*ta ✗	e*go*ís*ta ✓
egoístico	e*go-ís*ti*co ✗	e*go*ís*ti*co ✓

word	defult rule result	patched rules result
elastômero	e*las-tô*me*ro ✗	e*las*tô*me*ro ✓
elefantíase	e*le*fan*tí-a*se ✗	e*le*fan*tí*a*se ✓
eletrocardiógrafo	e*le*tro*car*di-ó*gra*fo ✗	e*le*tro*car*di*ó*gra*fo ✓
eletroímã	e*le*tro-í*mã ✗	e*le*tro*í*mã ✓
eletroquímica	e*le*tro-quí*mi*ca ✗	e*le*tro*quí*mi*ca ✓
elogiável	e*lo*gi-á*vel ✗	e*lo*gi*á*vel ✓
eluição	e*lu-i*ção ✗	e*lu*i*ção ✓
embainhar	em*ba-i*nhar ✗	em*ba*i*nhar ✓
embrião	em*bri-ão ✗	em*bri*ão ✓
ensaísmo	en*sa-ís*mo ✗	en*sa*ís*mo ✓
ensaísta	en*sa-ís*ta ✗	en*sa*ís*ta ✓
entusiástico	en*tu*si-ás*ti*co ✗	en*tu*si*ás*ti*co ✓
eólico	e-ó*li*co ✗	e*ó*li*co ✓
éon	é-on ✗	é*on ✓
equânime	e-quâ*ni*me ✗	e*quâ*ni*me ✓
equívoco	e-quí*vo*co ✗	e*quí*vo*co ✓
ergonômico	er*go-nô*mi*co ✗	er*go*nô*mi*co ✓
escorpião	es*cor*pi-ão ✗	es*cor*pi*ão ✓
esfigmomanômetro	es*fig*mo*ma-nô*me*tro ✗	es*fig*mo*ma*nô*me*tro ✓
espião	es*pi-ão ✗	es*pi*ão ✓
esponsais	es*pon*sa-is ✗	es*pon*sa-is ✗
esquálido	es-quá*li*do ✗	es*quá*li*do ✓
esteárico	es*te-á*ri*co ✗	es*te*á*ri*co ✓
estereótipo	es*te*re-ó*ti*po ✗	es*te*re*ó*ti*po ✓
estômago	es-tô*ma*go ✗	es*tô*ma*go ✓
esvair	es*va-ir ✗	es*va*ir ✓
etíope	e*tí-o*pe ✗	e*tí*o*pe ✓
etnônimo	et-nô*ni*mo ✗	et*nô*ni*mo ✓
eucarió́tico	eu*ca*ri-ó*ti*co ✗	eu*ca*ri*ó*ti*co ✓
eurasiático	eu*ra*si-á*ti*co ✗	eu*ra*si*á*ti*co ✓
evoluído	e*vo*lu-í*do ✗	e*vo*lu*í*do ✓
evoluir	e*vo*lu-ir ✗	e*vo*lu*ir ✓
excluído	ex*clu-í*do ✗	ex*clu*í*do ✓
excluir	ex*clu-ir ✗	ex*clu*ir ✓
exéquias	e*xé*qui.as ✗	e*xé*qui.as ✗
exequível	e*xe-quí*vel ✗	e*xe*quí*vel ✓
extrair	ex*tra-ir ✗	ex*tra*ir ✓
faim	fa-im ✗	fa-im ✗
faísca	fa-ís*ca ✗	fa*ís*ca ✓
farisaísmo	fa*ri*sa-ís*mo ✗	fa*ri*sa*ís*mo ✓
faúlha	fa-ú*lha ✗	fa*ú*lha ✓
feérico	fe-é*ri*co ✗	fe*é*ri*co ✓
feiura	fei-u*ra ✗	fei*u*ra ✓
fenômeno	fe-nô*me*no ✗	fe*nô*me*no ✓
fiável	fi-á*vel ✗	fi*á*vel ✓
fideísmo	fi*de-ís*mo ✗	fi*de*ís*mo ✓
filarmônica	fi*lar-mô*ni*ca ✗	fi*lar*mô*ni*ca ✓
filarmônico	fi*lar-mô*ni*co ✗	fi*lar*mô*ni*co ✓
financiável	fi*nan*ci-á*vel ✗	fi*nan*ci*á*vel ✓
fisiólogo	fi*si-ó*lo*go ✗	fi*si*ó*lo*go ✓
fiúza	fi-ú*za ✗	fi*ú*za ✓
fluidez	flu-i*dez ✗	flu-i*dez ✗
fluídico	flu-í*di*co ✗	flu*í*di*co ✓
fluir	flu-ir ✗	flu*ir ✓
flúor	flú-or ✗	flú*or ✓
folião	fo*li-ão ✗	fo*li*ão ✓
formaldeído	for*mal*de-í*do ✗	for*mal*de*í*do ✓
fotoquímica	fo*to-quí*mi*ca ✗	fo*to*quí*mi*ca ✓
fotoquímico	fo*to-quí*mi*co ✗	fo*to*quí*mi*co ✓
freático	fre-á*ti*co ✗	fre*á*ti*co ✓
friável	fri-á*vel ✗	fri*á*vel ✓
fruição	fru-i*ção ✗	fru*i*ção ✓
fruir	fru-ir ✗	fru*ir ✓
fuinha	fu-i*nha ✗	fu*i*nha ✓
futevôlei	fu*te-vô*lei ✗	fu*te*vô*lei ✓

word	defult rule result	patched rules result
gabião	ga*bi-ão ✗	ga*bi*ão ✓
gaélico	ga-é*li*co ✗	ga*é*li*co ✓
galeão	ga*le-ão ✗	ga*le*ão ✓
gastrointestinal	gas*tro-in*tes*ti*nal ✗	gas*tro*in*tes*ti*nal ✓
gastronômico	gas*tro-nô*mi*co ✗	gas*tro*nô*mi*co ✓
gaúcho	ga-ú*cho ✗	ga*ú*cho ✓
gavião	ga*vi-ão ✗	ga*vi*ão ✓
genuíno	ge*nu-í*no ✗	ge*nu*í*no ✓
geógrafo	ge-ó*gra*fo ✗	ge*ó*gra*fo ✓
geólogo	ge-ó*lo*go ✗	ge*ó*lo*go ✓
gigolô	gi*go-lô ✗	gi*go*lô ✓
gladíolo	gla*dí-o*lo ✗	gla*dí*o*lo ✓
glicoproteína	gli*co*pro*te-í*na ✗	gli*co*pro*te*í*na ✓
gnaisse	g.nais*se ✗	g.nais*se ✗
gnomo	g.no*mo ✗	gno*mo ✓
gnose	g.no*se ✗	gno*se ✓
gnosticismo	g.nos*ti*cis*mo ✗	gnos*ti*cis*mo ✓
gnóstico	g.nós*ti*co ✗	gnós*ti*co ✓
grafomaníaco	gra*fo*ma*ní-a*co ✗	gra*fo*ma*ní*a*co ✓
grapiúna	gra*pi-ú*na ✗	gra*pi*ú*na ✓
graúdo	gra-ú*do ✗	gra*ú*do ✓
graúna	gra-ú*na ✗	gra*ú*na ✓
guardião	guar*di-ão ✗	guar*di*ão ✓
guião	gui-ão ✗	gui*ão ✓
hagiógrafo	ha*gi-ó*gra*fo ✗	ha*gi*ó*gra*fo ✓
hanseático	han*se-á*ti*co ✗	han*se*á*ti*co ✓
hanseníase	han*se*ní-a*se ✗	han*se*ní*a*se ✓
harmônica	har-mô*ni*ca ✗	har*mô*ni*ca ✓
harmônico	har-mô*ni*co ✗	har*mô*ni*co ✓
hebraísmo	he*bra-ís*mo ✗	he*bra*ís*mo ✓
hebraísta	he*bra-ís*ta ✗	he*bra*ís*ta ✓
hegemônico	he*ge-mô*ni*co ✗	he*ge*mô*ni*co ✓
helíaco	he*lí-a*co ✗	he*lí*a*co ✓
heliógrafo	he*li-ó*gra*fo ✗	he*li*ó*gra*fo ✓
hemodiálise	he*mo*di-á*li*se ✗	he*mo*di*á*li*se ✓
heptacampeão	hep*ta*cam*pe-ão ✗	hep*ta*cam*pe*ão ✓
heroína	he*ro-í*na ✗	he*ro*í*na ✓
heroísmo	he*ro-ís*mo ✗	he*ro*ís*mo ✓
heteroátomo	he*te*ro-á*to*mo ✗	he*te*ro*á*to*mo ✓
heterônimo	he*te-rô*ni*mo ✗	he*te*rô*ni*mo ✓
hexacampeão	he*xa*cam*pe-ão ✗	he*xa*cam*pe*ão ✓
hidroavião	hi*dro*a*vi-ão ✗	hi*dro*a*vi*ão ✓
hidropônico	hi*dro-pô*ni*co ✗	hi*dro*pô*ni*co ✓
hinduísmo	hin*du-ís*mo ✗	hin*du*ís*mo ✓
hipocondríaco	hi*po*con*drí-a*co ✗	hi*po*con*drí*a*co ✓
hipônimo	hi-pô*ni*mo ✗	hi*pô*ni*mo ✓
histoquímica	his*to-quí*mi*ca ✗	his*to*quí*mi*ca ✓
homônimo	ho-mô*ni*mo ✗	ho*mô*ni*mo ✓
ictiólogo	ic*ti-ó*lo*go ✗	ic*ti*ó*lo*go ✓
ideólogo	i*de-ó*lo*go ✗	i*de*ó*lo*go ✓
iídiche	i-í*di*che ✗	i*í*di*che ✓
ilíaco	i*lí-a*co ✗	i*lí*a*co ✓
ilíada	i*lí-a*da ✗	i*lí*a*da ✓
imbuído	im*bu-í*do ✗	im*bu*í*do ✓
imbuir	im*bu-ir ✗	im*bu*ir ✓
imperdoável	im*per*do-á*vel ✗	im*per*do*á*vel ✓
impronunciável	im*pro*nun*ci-á*vel ✗	im*pro*nun*ci*á*vel ✓
inadiável	i*na*di-á*vel ✗	i*na*di*á*vel ✓
incluir	in*clu-ir ✗	in*clu*ir ✓
incômodo	in-cô*mo*do ✗	in*cô*mo*do ✓
inconciliável	in*con*ci*li-á*vel ✗	in*con*ci*li*á*vel ✓
indeiscente	in*de-is*cen*te ✗	in*de-is*cen*te ✗
inegociável	i*ne*go*ci-á*vel ✗	i*ne*go*ci*á*vel ✓
inequívoco	i*ne-quí*vo*co ✗	i*ne*quí*vo*co ✓
inexequível	i*ne*xe-quí*vel ✗	i*ne*xe*quí*vel ✓

word	defult rule result	patched rules result
influenciável	in*flu*en*ci-á*vel ✗	in*flu*en*ci*á*vel ✓
influído	in*flu-í*do ✗	in*flu*í*do ✓
influir	in*flu-ir ✗	in*flu*ir ✓
ingenuidade	in*ge*nu-i*da*de ✗	in*ge*nu*i*da*de ✓
iniciático	i*ni*ci-á*ti*co ✗	i*ni*ci*á*ti*co ✓
insaciável	in*sa*ci-á*vel ✗	in*sa*ci*á*vel ✓
instituído	ins*ti*tu-í*do ✗	ins*ti*tu*í*do ✓
instituir	ins*ti*tu-ir ✗	ins*ti*tu*ir ✓
instruído	ins*tru-í*do ✗	ins*tru*í*do ✓
instruir	ins*tru-ir ✗	ins*tru*ir ✓
insubstituível	in*subs*ti*tu-í*vel ✗	in*subs*ti*tu*í*vel ✓
intercambiável	in*ter*cam*bi-á*vel ✗	in*ter*cam*bi*á*vel ✓
interoceânico	in*te*ro*ce-â*ni*co ✗	in*te*ro*ce*â*ni*co ✓
intuição	in*tu-i*ção ✗	in*tu*i*ção ✓
intuir	in*tu-ir ✗	in*tu*ir ✓
intuitivo	in*tu-i*ti*vo ✗	in*tu-i*ti*vo ✗
invariável	in*va*ri-á*vel ✗	in*va*ri*á*vel ✓
inviável	in*vi-á*vel ✗	in*vi*á*vel ✓
iódico	i-ó*di*co ✗	i*ó*di*co ✓
irônico	i-rô*ni*co ✗	i*rô*ni*co ✓
iroquês	i*ro-quês ✗	i*ro*quês ✓
irreconciliável	ir*re*con*ci*li-á*vel ✗	ir*re*con*ci*li*á*vel ✓
irrefreável	ir*re*fre-á*vel ✗	ir*re*fre*á*vel ✓
irremediável	ir*re*me*di-á*vel ✗	ir*re*me*di*á*vel ✓
irrenunciável	ir*re*nun*ci-á*vel ✗	ir*re*nun*ci*á*vel ✓
irretorquível	ir*re*tor-quí*vel ✗	ir*re*tor*quí*vel ✓
isotônico	i*so-tô*ni*co ✗	i*so*tô*ni*co ✓
jabô	ja-bô ✗	ja*bô ✓
jesuíta	je*su-í*ta ✗	je*su*í*ta ✓
jesuítico	je*su-í*ti*co ✗	je*su*í*ti*co ✓
jesuitismo	je*su-i*tis*mo ✗	je*su*i*tis*mo ✓
joinvilense	jo-in*vi*len*se ✗	jo*in*vi*len*se ✓
judaísmo	ju*da-ís*mo ✗	ju*da*ís*mo ✓
judô	ju-dô ✗	ju*dô ✓
juizado	ju-i*za*do ✗	ju*i*za*do ✓
juíza	ju-í*za ✗	ju*í*za ✓
juiz	ju-iz ✗	ju*iz ✓
juízo	ju-í*zo ✗	ju*í*zo ✓
kuwaitiano	ku-wai*ti*a*no ✗	ku-wai*ti*a*no ✗
lacônico	la-cô*ni*co ✗	la*cô*ni*co ✓
ladainha	la*da-i*nha ✗	la*da*i*nha ✓
lamaísmo	la*ma-ís*mo ✗	la*ma*ís*mo ✓
lampião	lam*pi-ão ✗	lam*pi*ão ✓
laquê	la-quê ✗	la*quê ✓
leão	le-ão ✗	le*ão ✓
legião	le*gi-ão ✗	le*gi*ão ✓
litíase	li*tí-a*se ✗	li*tí*a*se ✓
luís	lu-ís ✗	lu*ís ✓
lusíada	lu*sí-a*da ✗	lu*sí*a*da ✓
luteína	lu*te-í*na ✗	lu*te*í*na ✓
macaúba	ma*ca-ú*ba ✗	ma*ca*ú*ba ✓
maçônico	ma-çô*ni*co ✗	ma*çô*ni*co ✓
macrobiótica	ma*cro*bi-ó*ti*ca ✗	ma*cro*bi*ó*ti*ca ✓
macrobiótico	ma*cro*bi-ó*ti*co ✗	ma*cro*bi*ó*ti*co ✓
macroeconômico	ma*cro*e*co-nô*mi*co ✗	ma*cro*e*co*nô*mi*co ✓
macrorregião	ma*cror*re*gi-ão ✗	ma*cror*re*gi*ão ✓
maiólica	mai-ó*li*ca ✗	mai*ó*li*ca ✓
maís	ma-ís ✗	ma*ís ✓
maiúscula	mai-ús*cu*la ✗	mai*ús*cu*la ✓
maiúsculo	mai-ús*cu*lo ✗	mai*ús*cu*lo ✓
malauiano	ma*lau-i*a*no ✗	ma*lau*i*a*no ✓
maleável	ma*le-á*vel ✗	ma*le*á*vel ✓
malthusianismo	mal-thu*si*a*nis*mo ✗	mal-thu*si*a*nis*mo ✗
malthusiano	mal-thu*si*a*no ✗	mal-thu*si*a*no ✗
mandrião	man*dri-ão ✗	man*dri*ão ✓

word	defult rule result	patched rules result
maníaco	ma*ní-a*co ✗	ma*ní*a*co ✓
maniqueísmo	ma*ni*que-ís*mo ✗	ma*ni*que*ís*mo ✓
maniqueísta	ma*ni*que-ís*ta ✗	ma*ni*que*ís*ta ✓
manômetro	ma-nô*me*tro ✗	ma*nô*me*tro ✓
mantô	man-tô ✗	man*tô ✓
marquês	mar-quês ✗	mar*quês ✓
maruim	ma*ru-im ✗	ma*ru-im ✗
masdeísmo	mas*de-ís*mo ✗	mas*de*ís*mo ✓
meão	me-ão ✗	me*ão ✓
mediático	me*di-á*ti*co ✗	me*di*á*ti*co ✓
medíocre	me*dí-o*cre ✗	me*dí*o*cre ✓
megalomaníaco	me*ga*lo*ma*ní-a*co ✗	me*ga*lo*ma*ní*a*co ✓
megawatt	me*ga-watt ✗	me*ga-watt ✗
meteórico	me*te-ó*ri*co ✗	me*te*ó*ri*co ✓
microbiólogo	mi*cro*bi-ó*lo*go ✗	mi*cro*bi*ó*lo*go ✓
micuim	mi*cu-im ✗	mi*cu-im ✗
midríase	mi*drí-a*se ✗	mi*drí*a*se ✓
mimeógrafo	mi*me-ó*gra*fo ✗	mi*me*ó*gra*fo ✓
míope	mí-o*pe ✗	mí*o*pe ✓
miríade	mi*rí-a*de ✗	mi*rí*a*de ✓
miúdo	mi-ú*do ✗	mi*ú*do ✓
mnemônica	m.ne-mô*ni*ca ✗	mne*mô*ni*ca ✓
mnemônico	m.ne-mô*ni*co ✗	mne*mô*ni*co ✓
moído	mo-í*do ✗	mo*í*do ✓
moinha	mo-i*nha ✗	mo*i*nha ✓
moinho	mo-i*nho ✗	mo*i*nho ✓
monomaníaco	mo*no*ma*ní-a*co ✗	mo*no*ma*ní*a*co ✓
monômero	mo-nô*me*ro ✗	mo*nô*me*ro ✓
monoteísmo	mo*no*te-ís*mo ✗	mo*no*te*ís*mo ✓
monoteísta	mo*no*te-ís*ta ✗	mo*no*te*ís*ta ✓
moquém	mo-quém ✗	mo*quém ✓
morrião	mor*ri-ão ✗	mor*ri*ão ✓
mucuim	mu*cu-im ✗	mu*cu-im ✗
museólogo	mu*se-ó*lo*go ✗	mu*se*ó*lo*go ✓
napoleão	na*po*le-ão ✗	na*po*le*ão ✓
neerlandês	ne*er*lan*dês ✓	ne*er*lan*dês ✓
neocolonialismo	ne*o*co*lo*ni*a*lis*mo ✓	ne-o*co*lo*ni*a*lis*mo ✗
neofascista	ne*o*fas*cis*ta ✓	ne-o*fas*cis*ta ✗
neófito	ne-ó*fi*to ✗	ne*ó*fi*to ✓
neolatino	ne*o*la*ti*no ✓	ne-o*la*ti*no ✗
neoliberalismo	ne*o*li*be*ra*lis*mo ✓	ne-o*li*be*ra*lis*mo ✗
neoliberal	ne*o*li*be*ral ✓	ne-o*li*be*ral ✗
neolítico	ne*o*lí*ti*co ✓	ne-o*lí*ti*co ✗
neologismo	ne*o*lo*gis*mo ✓	ne-o*lo*gis*mo ✗
neonato	ne*o*na*to ✓	ne-o*na*to ✗
neonazismo	ne*o*na*zis*mo ✓	ne-o*na*zis*mo ✗
neonazista	ne*o*na*zis*ta ✓	ne-o*na*zis*ta ✗
neon	ne*on ✓	ne-on ✗
néon	né-on ✗	né*on ✓
neoplatonismo	ne*o*pla*to*nis*mo ✓	ne-o*pla*to*nis*mo ✗
neozelandês	ne*o*ze*lan*dês ✓	ne-o*ze*lan*dês ✗
neurocirurgião	neu*ro*ci*rur*gi-ão ✗	neu*ro*ci*rur*gi*ão ✓
ninfomaníaca	nin*fo*ma*ní-a*ca ✗	nin*fo*ma*ní*a*ca ✓
ninfomaníaco	nin*fo*ma*ní-a*co ✗	nin*fo*ma*ní*a*co ✓
nobiliárquico	no*bi*li-ár*qui*co ✗	no*bi*li*ár*qui*co ✓
núpcias	núp*ci.as ✗	núp*ci.as ✗
oboísta	o*bo-ís*ta ✗	o*bo*ís*ta ✓
obstruído	obs*tru-í*do ✗	obs*tru*í*do ✓
obstruir	obs*tru-ir ✗	obs*tru*ir ✓
ocasião	o*ca*si-ão ✗	o*ca*si*ão ✓
oceânico	o*ce-â*ni*co ✗	o*ce*â*ni*co ✓
ocluir	o*clu-ir ✗	o*clu*ir ✓
octaédrico	oc*ta-é*dri*co ✗	oc*ta*é*dri*co ✓
odeão	o*de-ão ✗	o*de*ão ✓
odômetro	o-dô*me*tro ✗	o*dô*me*tro ✓

word	defult rule result	patched rules result
oleícola	o*le-í*co*la ✗	o*le*í*co*la ✓
oleína	o*le-í*na ✗	o*le*í*na ✓
olimpíada	o*lim*pí-a*da ✗	o*lim*pí*a*da ✓
oócito	o-ó*ci*to ✗	o*ó*ci*to ✓
opinião	o*pi*ni-ão ✗	o*pi*ni*ão ✓
orfeão	or*fe-ão ✗	or*fe*ão ✓
oriundo	o*ri-un*do ✗	o*ri*un*do ✓
ortodôntico	or*to-dôn*ti*co ✗	or*to*dôn*ti*co ✓
osteíte	os*te-í*te ✗	os*te*í*te ✓
painço	pa-in*ço ✗	pa-in*ço ✗
país	pa-ís ✗	pa*ís ✓
paleógrafo	pa*le-ó*gra*fo ✗	pa*le*ó*gra*fo ✓
pantagruélico	pan*ta*gru-é*li*co ✗	pan*ta*gru*é*li*co ✓
panteão	pan*te-ão ✗	pan*te*ão ✓
panteísmo	pan*te-ís*mo ✗	pan*te*ís*mo ✓
panteísta	pan*te-ís*ta ✗	pan*te*ís*ta ✓
paquímetro	pa-quí*me*tro ✗	pa*quí*me*tro ✓
paradisíaco	pa*ra*di*sí-a*co ✗	pa*ra*di*sí*a*co ✓
paraibano	pa*ra-i*ba*no ✗	pa*ra-i*ba*no ✗
paraíba	pa*ra-í*ba ✗	pa*ra*í*ba ✓
paraíso	pa*ra-í*so ✗	pa*ra*í*so ✓
parnaíba	par*na-í*ba ✗	par*na*í*ba ✓
parquê	par-quê ✗	par*quê ✓
parquímetro	par-quí*me*tro ✗	par*quí*me*tro ✓
parvoíce	par*vo-í*ce ✗	par*vo*í*ce ✓
patchuli	pat*chu*li ✓	pat-chu*li ✗
patenteável	pa*ten*te-á*vel ✗	pa*ten*te*á*vel ✓
patriótico	pa*tri-ó*ti*co ✗	pa*tri*ó*ti*co ✓
paul	pa-ul ✗	pa-ul ✗
paxiúba	pa*xi-ú*ba ✗	pa*xi*ú*ba ✓
peão	pe-ão ✗	pe*ão ✓
pecíolo	pe*cí-o*lo ✗	pe*cí*o*lo ✓
pediátrico	pe*di-á*tri*co ✗	pe*di*á*tri*co ✓
perdoável	per*do-á*vel ✗	per*do*á*vel ✓
periódico	pe*ri-ó*di*co ✗	pe*ri*ó*di*co ✓
período	pe*rí-o*do ✗	pe*rí*o*do ✓
periurbano	pe*ri-ur*ba*no ✗	pe*ri*ur*ba*no ✓
permeável	per*me-á*vel ✗	per*me*á*vel ✓
petroquímica	pe*tro-quí*mi*ca ✗	pe*tro*quí*mi*ca ✓
petroquímico	pe*tro-quí*mi*co ✗	pe*tro*quí*mi*co ✓
piauiense	pi*au-i*en*se ✗	pi*au*i*en*se ✓
picuinha	pi*cu-i*nha ✗	pi*cu*i*nha ✓
pindaíba	pin*da-í*ba ✗	pin*da*í*ba ✓
pintainho	pin*ta-i*nho ✗	pin*ta*i*nho ✓
piromaníaco	pi*ro*ma*ní-a*co ✗	pi*ro*ma*ní*a*co ✓
pitiríase	pi*ti*rí-a*se ✗	pi*ti*rí*a*se ✓
pium	pi-um ✗	pi*um ✓
pivô	pi-vô ✗	pi*vô ✓
pixaim	pi*xa-im ✗	pi*xa-im ✗
platônico	pla-tô*ni*co ✗	pla*tô*ni*co ✓
platô	pla-tô ✗	pla*tô ✓
plebeísmo	ple*be-ís*mo ✗	ple*be*ís*mo ✓
pneumática	p.neu*má*ti*ca ✗	pneu*má*ti*ca ✓
pneumático	p.neu*má*ti*co ✗	pneu*má*ti*co ✓
pneumococo	p.neu*mo*co*co ✗	pneu*mo*co*co ✓
pneumoconiose	p.neu*mo*co*ni*o*se ✗	pneu*mo*co*ni*o*se ✓
pneumogástrico	p.neu*mo*gás*tri*co ✗	pneu*mo*gás*tri*co ✓
pneumônico	p.neu-mô*ni*co ✗	pneu*mô*ni*co ✓
pneumotórax	p.neu*mo*tó*rax ✗	pneu*mo*tó*rax ✓
pneu	p.neu ✗	pneu ✓
poética	po-é*ti*ca ✗	po*é*ti*ca ✓
poético	po-é*ti*co ✗	po*é*ti*co ✓
poliéster	po*li-és*ter ✗	po*li*és*ter ✓
polifônico	po*li-fô*ni*co ✗	po*li*fô*ni*co ✓
politeísmo	po*li*te-ís*mo ✗	po*li*te*ís*mo ✓

word	defult rule result	patched rules result
politeísta	po*li*te-ís*ta ✗	po*li*te*ís*ta ✓
poluir	po*lu-ir ✗	po*lu*ir ✓
poraquê	po*ra-quê ✗	po*ra*quê ✓
pornô	por-nô ✗	por*nô ✓
porquê	por-quê ✗	por*quê ✓
possuído	pos*su-í*do ✗	pos*su*í*do ✓
possuidor	pos*su-i*dor ✗	pos*su*i*dor ✓
possuir	pos*su-ir ✗	pos*su*ir ✓
preâmbulo	pre-âm*bu*lo ✗	pre*âm*bu*lo ✓
preênsil	pre-ên*sil ✗	pre*ên*sil ✓
prejuízo	pre*ju-í*zo ✗	pre*ju*í*zo ✓
presidenciável	pre*si*den*ci-á*vel ✗	pre*si*den*ci*á*vel ✓
procaína	pro*ca-í*na ✗	pro*ca*í*na ✓
procariótico	pro*ca*ri-ó*ti*co ✗	pro*ca*ri*ó*ti*co ✓
prolegômenos	pro*le-gô*me*nos ✗	pro*le*gô*me*nos ✓
prosaísmo	pro*sa-ís*mo ✗	pro*sa*ís*mo ✓
prostituição	pros*ti*tu-i*ção ✗	pros*ti*tu*i*ção ✓
prostituir	pros*ti*tu-ir ✗	pros*ti*tu*ir ✓
proteína	pro*te-í*na ✗	pro*te*í*na ✓
proteólise	pro*te-ó*li*se ✗	pro*te*ó*li*se ✓
pseudônimo	p.seu-dô*ni*mo ✗	pseu*dô*ni*mo ✓
psicanálise	p.si*ca*ná*li*se ✗	psi*ca*ná*li*se ✓
psicanalista	p.si*ca*na*lis*ta ✗	psi*ca*na*lis*ta ✓
psicanalítico	p.si*ca*na*lí*ti*co ✗	psi*ca*na*lí*ti*co ✓
psicodrama	p.si*co*dra*ma ✗	psi*co*dra*ma ✓
psicógrafo	p.si*có*gra*fo ✗	psi*có*gra*fo ✓
psicolinguista	p.si*co*lin*guis*ta ✗	psi*co*lin*guis*ta ✓
psicologicamente	p.si*co*lo*gi*ca*men*te ✗	psi*co*lo*gi*ca*men*te ✓
psicológico	p.si*co*ló*gi*co ✗	psi*co*ló*gi*co ✓
psicologismo	p.si*co*lo*gis*mo ✗	psi*co*lo*gis*mo ✓
psicologista	p.si*co*lo*gis*ta ✗	psi*co*lo*gis*ta ✓
psicologizar	p.si*co*lo*gi*zar ✗	psi*co*lo*gi*zar ✓
psicólogo	p.si*có*lo*go ✗	psi*có*lo*go ✓
psicométrico	p.si*co*mé*tri*co ✗	psi*co*mé*tri*co ✓
psicomotor	p.si*co*mo*tor ✗	psi*co*mo*tor ✓
psicomotricidade	p.si*co*mo*tri*ci*da*de ✗	psi*co*mo*tri*ci*da*de ✓
psicopata	p.si*co*pa*ta ✗	psi*co*pa*ta ✓
psicopático	p.si*co*pá*ti*co ✗	psi*co*pá*ti*co ✓
psicopatológico	p.si*co*pa*to*ló*gi*co ✗	psi*co*pa*to*ló*gi*co ✓
psicopedagógico	p.si*co*pe*da*gó*gi*co ✗	psi*co*pe*da*gó*gi*co ✓
psicopedagogo	p.si*co*pe*da*go*go ✗	psi*co*pe*da*go*go ✓
psicose	p.si*co*se ✗	psi*co*se ✓
psicossexual	p.si*cos*se*xu*al ✗	psi*cos*se*xu*al ✓
psicossomático	p.si*cos*so*má*ti*co ✗	psi*cos*so*má*ti*co ✓
psicotécnico	p.si*co*téc*ni*co ✗	psi*co*téc*ni*co ✓
psicoterapeuta	p.si*co*te*ra*peu*ta ✗	psi*co*te*ra*peu*ta ✓
psicoterápico	p.si*co*te*rá*pi*co ✗	psi*co*te*rá*pi*co ✓
psicótico	p.si*có*ti*co ✗	psi*có*ti*co ✓
psicotrópico	p.si*co*tró*pi*co ✗	psi*co*tró*pi*co ✓
psique	p.si*que ✗	psi*que ✓
psiquiatra	p.si*qui*a*tra ✗	psi*qui*a*tra ✓
psiquiátrico	p.si*qui-á*tri*co ✗	psi*qui*á*tri*co ✓
psíquico	p.sí*qui*co ✗	psí*qui*co ✓
psiquismo	p.si*quis*mo ✗	psi*quis*mo ✓
psitaciforme	p.si*ta*ci*for*me ✗	psi*ta*ci*for*me ✓
psitacismo	p.si*ta*cis*mo ✗	psi*ta*cis*mo ✓
psoas	p.so*as ✗	p.so*as ✗
psoríase	p.so*rí-a*se ✗	p.so*rí*a*se ✗
pterossauro	p.te*ros*sau*ro ✗	pte*ros*sau*ro ✓
ptose	p.to*se ✗	pto*se ✓
pulôver	pu-lô*ver ✗	pu*lô*ver ✓
quartzito	quart*zi*to ✓	quart*zi*to ✓
quartzo	quart*zo ✓	quart*zo ✓
questiúncula	ques*ti-ún*cu*la ✗	ques*ti*ún*cu*la ✓
quilômetro	qui-lô*me*tro ✗	qui*lô*me*tro ✓

word	defult rule result	patched rules result
quilowatt	qui*lo-watt ✗	qui*lo-watt ✗
quiproquó	qui*pro-quó ✗	qui*pro*quó ✓
radiólise	ra*di-ó*li*se ✗	ra*di*ó*li*se ✓
rainha	ra-i*nha ✗	ra*i*nha ✓
raiz	ra-iz ✗	ra*iz ✓
randômico	ran-dô*mi*co ✗	ran*dô*mi*co ✓
raquítico	ra-quí*ti*co ✗	ra*quí*ti*co ✓
ravióli	ra*vi-ó*li ✗	ra*vi*ó*li ✓
razoável	ra*zo-á*vel ✗	ra*zo*á*vel ✓
rebelião	re*be*li-ão ✗	re*be*li*ão ✓
recaída	re*ca-í*da ✗	re*ca*í*da ✓
recair	re*ca-ir ✗	re*ca*ir ✓
recôncavo	re-côn*ca*vo ✗	re*côn*ca*vo ✓
recôndito	re-côn*di*to ✗	re*côn*di*to ✓
reconstituição	re*cons*ti*tu-i*ção ✗	re*cons*ti*tu*i*ção ✓
reconstituinte	re*cons*ti*tu-in*te ✗	re*cons*ti*tu*in*te ✓
reconstituir	re*cons*ti*tu-ir ✗	re*cons*ti*tu*ir ✓
reconstruído	re*cons*tru-í*do ✗	re*cons*tru*í*do ✓
reconstruir	re*cons*tru-ir ✗	re*cons*tru*ir ✓
redistribuir	re*dis*tri*bu-ir ✗	re*dis*tri*bu*ir ✓
refluir	re*flu-ir ✗	re*flu*ir ✓
região	re*gi-ão ✗	re*gi*ão ✓
reimplantação	re-im*plan*ta*ção ✗	re*im*plan*ta*ção ✓
reimplantar	re-im*plan*tar ✗	re*im*plan*tar ✓
reimportar	re-im*por*tar ✗	re*im*por*tar ✓
reimpressão	re-im*pres*são ✗	re*im*pres*são ✓
reimprimir	re-im*pri*mir ✗	re*im*pri*mir ✓
reincidente	re-in*ci*den*te ✗	re*in*ci*den*te ✓
reincidir	re-in*ci*dir ✗	re*in*ci*dir ✓
reincorporação	re-in*cor*po*ra*ção ✗	re*in*cor*po*ra*ção ✓
reindexação	re-in*de*xa*ção ✗	re-in*de*xa*ção ✗
reingressar	re-in*gres*sar ✗	re*in*gres*sar ✓
reingresso	re-in*gres*so ✗	re*in*gres*so ✓
reinscrever	re-ins*cre*ver ✗	re*ins*cre*ver ✓
reinserção	re-in*ser*ção ✗	re*in*ser*ção ✓
reinserir	re-in*se*rir ✗	re*in*se*rir ✓
reinstalação	re-ins*ta*la*ção ✗	re*ins*ta*la*ção ✓
reinstalar	re-ins*ta*lar ✗	re*ins*ta*lar ✓
reinstituir	re-ins*ti*tu-ir ✗	re*ins*ti*tu*ir ✓
reintegração	re-in*te*gra*ção ✗	re*in*te*gra*ção ✓
reintegrar	re-in*te*grar ✗	re*in*te*grar ✓
reintrodução	re-in*tro*du*ção ✗	re*in*tro*du*ção ✓
reintroduzir	re-in*tro*du*zir ✗	re*in*tro*du*zir ✓
reinventar	re-in*ven*tar ✗	re*in*ven*tar ✓
reinvestir	re-in*ves*tir ✗	re*in*ves*tir ✓
religião	re*li*gi-ão ✗	re*li*gi*ão ✓
restituição	res*ti*tu-i*ção ✗	res*ti*tu*i*ção ✓
restituir	res*ti*tu-ir ✗	res*ti*tu*ir ✓
retraído	re*tra-í*do ✗	re*tra*í*do ✓
retrair	re*tra-ir ✗	re*tra*ir ✓
retribuição	re*tri*bu-i*ção ✗	re*tri*bu*i*ção ✓
retribuir	re*tri*bu-ir ✗	re*tri*bu*ir ✓
reunir	re-u*nir ✗	re-u*nir ✗
reurbanizar	re-ur*ba*ni*zar ✗	re-ur*ba*ni*zar ✗
robô	ro-bô ✗	ro*bô ✓
roído	ro-í*do ✗	ro*í*do ✓
roséola	ro*sé-o*la ✗	ro*sé*o*la ✓
ruão	ru-ão ✗	ru*ão ✓
rubéola	ru*bé-o*la ✗	ru*bé*o*la ✓
rufião	ru*fi-ão ✗	ru*fi*ão ✓
ruído	ru-í*do ✗	ru*í*do ✓
ruim	ru-im ✗	ru-im ✗
ruína	ru-í*na ✗	ru*í*na ✓
ruindade	ru-in*da*de ✗	ru*in*da*de ✓
ruinoso	ru-i*no*so ✗	ru*i*no*so ✓

word	defult rule result	patched rules result
ruir	ru-ir ✗	ru*ir ✓
rutherford	ru-ther*ford ✗	ru-ther*ford ✗
saída	sa-í*da ✗	sa*í*da ✓
saído	sa-í*do ✗	sa*í*do ✓
sair	sa-ir ✗	sa*ir ✓
salomônico	sa*lo-mô*ni*co ✗	sa*lo*mô*ni*co ✓
sambaíba	sam*ba-í*ba ✗	sam*ba*í*ba ✓
sanduíche	san*du-í*che ✗	san*du*í*che ✓
saquê	sa-quê ✗	sa*quê ✓
sardônico	sar-dô*ni*co ✗	sar*dô*ni*co ✓
satiríase	sa*ti*rí-a*se ✗	sa*ti*rí*a*se ✓
saúde	sa-ú*de ✗	sa*ú*de ✓
saúva	sa-ú*va ✗	sa*ú*va ✓
sebastião	se*bas*ti-ão ✗	se*bas*ti*ão ✓
semiárido	se*mi-á*ri*do ✗	se*mi*á*ri*do ✓
siálico	si-á*li*co ✗	si*á*li*co ✓
sinfônica	sin-fô*ni*ca ✗	sin*fô*ni*ca ✓
sinfônico	sin-fô*ni*co ✗	sin*fô*ni*co ✓
sinônimo	si-nô*ni*mo ✗	si*nô*ni*mo ✓
siríaco	si*rí-a*co ✗	si*rí*a*co ✓
sobressair	so*bres*sa-ir ✗	so*bres*sa*ir ✓
sociólogo	so*ci-ó*lo*go ✗	so*ci*ó*lo*go ✓
sonômetro	so-nô*me*tro ✗	so*nô*me*tro ✓
soviético	so*vi-é*ti*co ✗	so*vi*é*ti*co ✓
stalinismo	s.ta*li*nis*mo ✗	sta*li*nis*mo ✓
stalinista	s.ta*li*nis*ta ✗	sta*li*nis*ta ✓
suaíli	su*a-í*li ✗	su*a*í*li ✓
suão	su-ão ✗	su*ão ✓
suástica	su-ás*ti*ca ✗	su*ás*ti*ca ✓
subaquático	su*ba-quá*ti*co ✗	su*ba*quá*ti*co ✓
subdiácono	sub*di-á*co*no ✗	sub*di*á*co*no ✓
subliminar	su.b-li*mi*nar ✗	su.b-li*mi*nar ✗
sublingual	su.b-lin*gual ✗	su.b-lin*gual ✗
sublinha	su.b-li*nha ✗	su.b-li*nha ✗
subliteratura	su.b-li*te*ra*tu*ra ✗	su.b-li*te*ra*tu*ra ✗
sublocação	su.b-lo*ca*ção ✗	sub*lo*ca*ção ✓
sublocar	su.b-lo*car ✗	sub*lo*car ✓
sublunar	su.b-lu*nar ✗	sub*lu*nar ✓
substituição	subs*ti*tu-i*ção ✗	subs*ti*tu*i*ção ✓
substituído	subs*ti*tu-í*do ✗	subs*ti*tu*í*do ✓
substituir	subs*ti*tu-ir ✗	subs*ti*tu*ir ✓
substituível	subs*ti*tu-í*vel ✗	subs*ti*tu*í*vel ✓
subtraído	sub*tra-í*do ✗	sub*tra*í*do ✓
subtrair	sub*tra-ir ✗	sub*tra*ir ✓
suéter	su-é*ter ✗	su*é*ter ✓
suíça	su-í*ça ✗	su*í*ça ✓
suíço	su-í*ço ✗	su*í*ço ✓
suingue	su-in*gue ✗	su*in*gue ✓
suíno	su-í*no ✗	su*í*no ✓
suíte	su-í*te ✗	su*í*te ✓
sumô	su-mô ✗	su*mô ✓
supercampeão	su*per*cam*pe-ão ✗	su*per*cam*pe*ão ✓
supersônico	su*per-sô*ni*co ✗	su*per*sô*ni*co ✓
tabelião	ta*be*li-ão ✗	ta*be*li*ão ✓
tabuão	ta*bu-ão ✗	ta*bu*ão ✓
tacômetro	ta-cô*me*tro ✗	ta*cô*me*tro ✓
tainha	ta-i*nha ✗	ta*i*nha ✓
taiwanês	tai-wa*nês ✗	tai-wa*nês ✗
talião	ta*li-ão ✗	ta*li*ão ✓
taquígrafo	ta-quí*gra*fo ✗	ta*quí*gra*fo ✓
taquímetro	ta-quí*me*tro ✗	ta*quí*me*tro ✓
tarô	ta-rô ✗	ta*rô ✓
tataravô	ta*ta*ra-vô ✗	ta*ta*ra*vô ✓
taxonômico	ta*xo-nô*mi*co ✗	ta*xo*nô*mi*co ✓
tcheco	t.che*co ✗	tche*co ✓

word	defult rule result	patched rules result
teísmo	te-ís*mo ✗	te*ís*mo ✓
teísta	te-ís*ta ✗	te*ís*ta ✓
telefônico	te*le-fô*ni*co ✗	te*le*fô*ni*co ✓
teníase	te*ní-a*se ✗	te*ní*a*se ✓
teófobo	te-ó*fo*bo ✗	te*ó*fo*bo ✓
teólogo	te-ó*lo*go ✗	te*ó*lo*go ✓
teônimo	te-ô*ni*mo ✗	te*ô*ni*mo ✓
teórica	te-ó*ri*ca ✗	te*ó*ri*ca ✓
teórico	te-ó*ri*co ✗	te*ó*ri*co ✓
termômetro	ter-mô*me*tro ✗	ter*mô*me*tro ✓
termoquímica	ter*mo-quí*mi*ca ✗	ter*mo*quí*mi*ca ✓
termoquímico	ter*mo-quí*mi*co ✗	ter*mo*quí*mi*co ✓
tetracampeão	te*tra*cam*pe-ão ✗	te*tra*cam*pe*ão ✓
tetraédrico	te*tra-é*dri*co ✗	te*tra*é*dri*co ✓
tetravô	te*tra-vô ✗	te*tra*vô ✓
teúrgico	te-úr*gi*co ✗	te*úr*gi*co ✓
teutônico	teu-tô*ni*co ✗	teu*tô*ni*co ✓
tmese	t.me*se ✗	tme*se ✓
toboágua	to*bo-á*gua ✗	to*bo*á*gua ✓
topônimo	to-pô*ni*mo ✗	to*pô*ni*mo ✓
torquês	tor-quês ✗	tor*quês ✓
torreão	tor*re-ão ✗	tor*re*ão ✓
toxicômano	to*xi-cô*ma*no ✗	to*xi*cô*ma*no ✓
tragediógrafo	tra*ge*di-ó*gra*fo ✗	tra*ge*di*ó*gra*fo ✓
traído	tra-í*do ✗	tra*í*do ✓
traíra	tra-í*ra ✗	tra*í*ra ✓
trair	tra-ir ✗	tra*ir ✓
transeunte	tran*se-un*te ✗	tran*se-un*te ✗
transoceânico	tran*so*ce-â*ni*co ✗	tran*so*ce*â*ni*co ✓
tríade	trí-a*de ✗	trí*a*de ✓
tricampeão	tri*cam*pe-ão ✗	tri*cam*pe*ão ✓
tricô	tri-cô ✗	tri*cô ✓
tripanossomíase	tri*pa*nos*so*mí-a*se ✗	tri*pa*nos*so*mí*a*se ✓
trisavô	tri*sa-vô ✗	tri*sa*vô ✓
triunfador	tri-un*fa*dor ✗	tri*un*fa*dor ✓
triunfalismo	tri-un*fa*lis*mo ✗	tri*un*fa*lis*mo ✓
triunfalista	tri-un*fa*lis*ta ✗	tri*un*fa*lis*ta ✓
triunfal	tri-un*fal ✗	tri*un*fal ✓
triunfante	tri-un*fan*te ✗	tri*un*fan*te ✓
triunfar	tri-un*far ✗	tri*un*far ✓
triunfo	tri-un*fo ✗	tri*un*fo ✓
triúnviro	tri-ún*vi*ro ✗	tri*ún*vi*ro ✓
truísmo	tru-ís*mo ✗	tru*ís*mo ✓
tuim	tu-im ✗	tu*im ✓
tzarista	t.za*ris*ta ✗	tza*ris*ta ✓
uísque	u-ís*que ✗	u*ís*que ✓
uíste	u-ís*te ✗	u*ís*te ✓
umbaúba	um*ba-ú*ba ✗	um*ba*ú*ba ✓
união	u*ni-ão ✗	u*ni*ão ✓
usucapião	u*su*ca*pi-ão ✗	u*su*ca*pi*ão ✓
usufruir	u*su*fru-ir ✗	u*su*fru*ir ✓
uveíte	u*ve-í*te ✗	u*ve*í*te ✓
vacúolo	va*cú-o*lo ✗	va*cú*o*lo ✓
variável	va*ri-á*vel ✗	va*ri*á*vel ✓
varíola	va*rí-o*la ✗	va*rí*o*la ✓
vascaíno	vas*ca-í*no ✗	vas*ca*í*no ✓
veículo	ve-í*cu*lo ✗	ve*í*cu*lo ✓
viável	vi-á*vel ✗	vi*á*vel ✓
vibrião	vi*bri-ão ✗	vi*bri*ão ✓
viés	vi-és ✗	vi*és ✓
vitriólico	vi*tri-ó*li*co ✗	vi*tri*ó*li*co ✓
viúva	vi-ú*va ✗	vi*ú*va ✓
viúvo	vi-ú*vo ✗	vi*ú*vo ✓
vovô	vo-vô ✗	vo*vô ✓
voyeurismo	voy-eu*ris*mo ✗	voy-eu*ris*mo ✗

word	defult rule result	patched rules result
voyeurista	voy-eu*ris*ta ✗	voy-eu*ris*ta ✗
voyeurístico	voy-eu*rís*ti*co ✗	voy-eu*rís*ti*co ✗
xilocaína	xi*lo*ca-í*na ✗	xi*lo*ca*í*na ✓
xintoísmo	xin*to-ís*mo ✗	xin*to*ís*mo ✓
xintoísta	xin*to-ís*ta ✗	xin*to*ís*ta ✓
zodíaco	zo*dí-a*co ✗	zo*dí*a*co ✓
zoósporo	zo-ós*po*ro ✗	zo*ós*po*ro ✓
total correct	17	811