

Enhancing TeX Hyphenation Rules for Portuguese

Leonardo Araujo and Aline Benevides

Abstract

Portuguese hyphenation rules have been available for more than 35 years and have performed well. Nonetheless, they still make mistakes and leave some hyphenation points unmarked. Although most undetected hyphenation points are located near word boundaries, which are irrelevant for TeX typographic purposes, they are still useful for hyphenating proper nouns, new words, or pseudowords, and for use in other applications, such as text-to-speech conversion. A list of 49,528 hyphenated words acquired from online dictionaries was used to analyze the default rules' hyphenation performance, and a set of 120 new rules is proposed to be added to the hyphenation rules for Portuguese. In addition, Patgen was used to create new rules from scratch or to improve the previous set of rules. Patgen's set of rules doesn't show good generalization capability. In the end, the set of handcrafted patched rules shows the best results.

1 Introduction

Automatic hyphenation is crucial part of document preparation systems, especially in typesetting and desktop publishing. It improves readability and aesthetics in text layout. It avoids the necessity of editors or typographers interventions to insert hyphens at suitable locations and it might be easily adjusted to different languages.

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 software (e.g. LibreOffice, Mozilla Firefox, Mozilla Thunderbird, Google Chrome, macOS, InDesign, memoQ, Opera, Affinity Publisher, among others [11]). Hunspell uses TeX hyphenation rules [12, 14], 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 [22], and offers the flexibility to create rules for any currently unsupported languages.

In this paper, we evaluate the current hyphenation rules initially proposed by de Rezende (1987) [7] and later updated by de Rezende and Almeida (2015) [8]. We defined our golden standard hyphenation dictionary based on six different Portuguese

online dictionaries, and used it to apprise the performance of different set of rules. Using grammatical rules and phonological principles, we proposed a set of rules to complement the existing standard hyphenation rules used in TeX, and we also created new set of rules using Patgen.

This paper is organized as follows: Section 2 discusses hyphenation as a general tool; Section 3 reviews the existing TeX hyphenation rules for Portuguese; Section 4 describes the creation of our hyphenation dictionary to be used as a references; Section 5 presents the results of applying the default TeX rules to this dictionary; Section 6 summarizes our proposed handcrafted patch rules to address errors found in the default rules; in Section 7 we use Patgen to create a set of rules from scratch, built on the default rules or built on the here proposed set of rules; finally, we conclude the paper in Section 8.

2 Hyphenation

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 [6], few documents used such sign until the end of the 18th century [5].

In some cases, hyphenation can hinder smooth reading and should be avoided. For example, in children's literature, it can disrupts the reading flow and comprehension. Similarly, in large fonts or headlines, hyphenation appears visually unappealing. In technical documents, it may cause confusion when applied to specialized terms. In opposition, large or small spaces between words can also impose difficulty on 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. In left-aligned text, consistent spacing can eliminate the risk of river formation, ensuring a consistent and aesthetically pleasing appearance of the text. Likewise, in justified text, hyphenation can create awkward spacing, reducing readability. Therefore, automatic hyphenation plays an important role in good typesetting.

Another important matter to consider is ambiguities that might arise 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* (for Portuguese, some examples are: *de-putada*, *fede-ração*, *acu-mula*, *após-tolo*, *cú-bico*). Hyphenations that might lead the reader to pronounce a word incorrectly should also be prevented (e.g. *considera-tion*, in English 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-chaе-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, the word *en-tres-sa-fra* is preferably split as *entres-safra* in opposition to *en-tressafra* or *entressa-fra*, and the word *fe-liz-men-te* is preferably split as *feliz-mente* in opposition to *fe-lizmente* or *felizmen-te*. In both cases, keeping the two morphemes together was the primary motivation. 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 preferable over splitting a double vowel).

Only the analysis of immediate surrounding may not be enough to determine a potential hyphenation point. For example, consider *dem-o-crat* and *democ-ra-cy*, where immediate surrounding of *e* does not suffice to determine the break location. Even after chosen patterns that makes hyphenation pretty straightforward, there might be exceptions. Consider the sequence *tion*. As it is commonly a suffix in English, for morphological reasons, it is preferable to always pause before this pattern, as in *celebra-tion*, *explana-tion* and *motiva-tion*. However, the word *cation* is hyphenated as *cat-ion*, making its etymology determinant to the way this word is split.

A word with different meanings may be hyphenated differently according to its 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).” [20]. In Portuguese, the word *sublinha* might be hyphenated in two different forms: *su-bli-nha* when representing the inflected form of verb *sublinhar* (underline as a verb) or *sub-li-nha* when referring to the line under (underline as a noun).

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 morphological constituents, creating a conflicting relation with the morphological preference described previously. Prefixes *bis-*, *des-* and *in-* are examples. The correct syllabifications are *bi-sa-vô*, *de-sem-bar-que*, *i-na-ti-vo* and *i-nobs-tan-te*¹, where the prefixes are split into two syllables.

Each language has its hyphenation rules, which can be categorized into four groups: morphology (etymology), focusing on meaningful word parts (prefixes, suffixes); phonology, driven by syllable breaks in speech; orthography, based on standard spelling conventions; and semantics, which considers context to avoid ambiguous or awkward splits. 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 might use a dictionary based approach, which will restrict the hyphenation possibilities to those entries in the dictionary, or a algorithmic base approach, which might be applied to whatever sequence found in a text. An algorithmic approach might use a logic system to analyse words and apply the hyphenation rules of a given language. A rule base model might include the recognition of prefixes, suffixes, morphemes, or some sequences suitable for the inclusion of a break. Another approach is the usage of patter 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. Finally, a mixed based model might also be used. It combines two or more of the previously described methods to enhance hyphenation accuracy and flexibility.

¹ The rule of syllable division could lead to two possible partitions, based on phonological or morphological forces: *i-nobs-tan-te* and *in-obs-tan-te*, but the first is preferable.

The original \TeX hyphenation algorithm was introduced by Knuth (1977) [13]. It was focused on the English language and used four main rules: (1) each part should contain a vowel other than final *e*; (2) suffix removal; (3) prefix removal; and (4) vowel-consonant-consonant-vowel (VCCV) breaking (but combine *h* with the previous letter if it is a consonant). Tests showed it could find 40% of the allowed hyphen locations [16]. The hyphenation algorithm proposed by Frank M. Liang and adopted in \TeX uses the notion of competing patterns [16, 17]. A database of hyphenated words is swept looking for hyphenating and inhibiting patterns. The algorithm introduced in \TeX 82 uses five alternating levels of hyphenating and inhibiting patterns. Patgen, the program for pattern generation based on a corpus, was created by Liang [15] and was used to create hyphenating patterns for many languages [26, 27, 30, 29, 24].

The effective hyphenation of words by \TeX will actually depend 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`, which defines the minimum sequence length of characters at the left and right borders before any hyphenation is allowed [28].

3 \TeX hyphenation rules for Portuguese

\TeX hyphenation rules for Portuguese have 307 rules in total [7, 8], being able to hyphenate correctly 92% of Portuguese words, when analyzing the hyphenation dictionary proposed here (see Section 4). However, by observing the erroneous or missing cases, we analyze here the default rules to identify areas for improvements.

Out of the default \TeX rules, 252 (82%) follow the pattern `1CV`, which represents the recurring CV syllables in Portuguese. As for the written consonant characters, 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 written vowel characters (*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 consonants². It is worth noting that there might be other rules or exceptions in the hyphenation patterns not covered

² In Portuguese, usually the letter *h* appears at word initial position or before the consonants *c, l, or n*. Therefore, it is not worthy encouraging a hyphenation point before an *hV*. Examples of words that have an *h* in the middle and not preceded by *c, l, or n* are: *Bahia, Corinthians, show, shopping, Matheus*, and *sushi*. They are proper names or loanwords whose spelling have not been adapted to Portuguese.

by these default rules. To accommodate exceptions, the following rules were proposed [7]:

- 20 rules created for cases involving consonants *b, c, d, f, g, k, p, t, v, or w*³ followed by *l* or *r*;
- 3 rules for *c, l* or *n* followed by *h*;
- 23 distinct patterns were introduced to indicate hyphenation points between vowels or between *c*'s, *r*'s, and *s*'s;
- 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;
- 1 pattern represented as `1-`, denoting that a hyphen indeed serves as a beneficial hyphenation point.

4 Creating the reference Portuguese hyphenation dictionary

Our set of Portuguese words was created from those in the CETENFolha corpus [18], augmented with the word list from Palavras NET [21] and from the Portuguese Wikipedia dump [33]. From the Wikipedia dump data, we narrowed down the selection to a subset of 50,721 words, accounting for 95% of occurrences in the Wikipedia dump (the initial set had 419,578 words). This threshold was set to filter out typos and infrequent words. Finally, 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 [19], Priberam [23], Wikcionário [32], Aulete [1], Portal da Língua Portuguesa [4], and Dicio [9]. This curation process resulted in a final dictionary containing 85,638 words. We have also manually performed corrections in 10 entries, we judged necessary.

The fact that the Portuguese language relies on phonology as a key factor for hyphenation requires an understanding of how stress influences this process. Every word in the language with two or more syllables has one more prominent syllable, which is considered the stressed syllable. The stress can fall on any of the last three syllables of the word, counted from the right margin. Words are classified as oxytone when the last syllable is stressed, paroxytone when the penultimate syllable is stressed, or proparoxytone when the third-to-last syllable is stressed. The stress pattern considered unmarked (the most common) does not receive a diacritical mark; the

³ We don't know why they have added the rules `1w21` and `1w2r`, since there is no Portuguese word with such sequence, we could only find foreignism like *showroom, wrestling, bowling*, or other rather rarer foreign words.

marked patterns, including all proparoxytones, do receive a diacritical mark. This marking serves to indicate to the reader that the stressed syllable follows a less common pattern. The use of diacritical marks to indicate stress is an important clue for determining how a word can be hyphenated. For example, the words *saúde* and *saída*, which have stress marked on the vowels *u* and *i*, respectively, clearly indicate that these vowels form their own syllables: *sa-ú-de* and *sa-í-da*. However, there are ambiguous cases, both for speakers and in linguistic literature, such as *his-tória* (or *his-tó-ri-a*) and *car-tó-rio* (or *car-tó-ri-o*). The question is whether final vowel clusters like *-ia* and *-io* are diphthongs or hiatuses. Normative grammar indicates they are paroxytones, but also acknowledges that they can be apparent proparoxytones [3, 2].

In the Portuguese Wikitionary we found 1,848 words marked as apparent proparoxytones. Apparent proparoxytones are words that appear to have stress on the antepenultimate syllable due to post-tonic vowel sequences treated as rising diphthongs. Despite their appearance, they follow the accentuation rules for paroxytones. Apparent proparoxytones can create confusion in syllabification and hyphenation because their stress pattern suggests different breaking points. For example, the word *história* could be hyphenated in two ways: *his-tó-ri-a* (proparoxytone) or *his-tó-ria* (paroxytone). This dual potential arises from treating the final vowel sequences as either separate syllables or as parts of diphthongs. In linguistic terms, accurate syllabification should consider the intended pronunciation and morphological structure, aligning with standard accentuation rules for the language [25]. For those words marked as apparent proparoxytones, we have accepted both hyphenations.

We have extracted 4 sub-dictionaries from the created hyphenation dictionary: the first dictionary is made of 15,842 words where the hyphenation given by all six dictionaries is the same; the next dictionary is made of 15,642 where five out of the six dictionaries agree with the same hyphenation; then we have a dictionary with 10,299 words with four agreed hyphenations, and finally a dictionary with 7,745 words where just three dictionaries agree on the hyphenation.

Those four dictionaries were used to gradually incorporate new rules to fix hyphenation issues in Portuguese, whether by analyzing and proposing fixes or by using Patgen to create a new set of rules.

5 Performance of the default rules

In this section, we evaluate the effectiveness of the default T_EX hyphenation rules for Portuguese by testing them against the comprehensive hyphenation dictionaries described in Section 4. This analysis helps us understand the accuracy and limitations of the existing rules and sets a baseline for further enhancements. The results of this evaluation are presented in Table 1.

Table 1: Results of T_EX default hyphenation rules on the dictionaries created from online dictionaries hyphenations.

word list	correct	incorrect	missing	entries
6 agrees	98.1%	0.2%	1.7%	15,842
5 agrees	89.4%	7.3%	3.8%	15,642
4 agrees	87.4%	8.6%	5.0%	10,299
3 agrees	93.9%	1.0%	5.4%	7,766
total	92.5%	4.3%	3.7%	49,549

Some words might have a bad and a missing hyphenation point concomitantly, therefore the percentage in each line might add up to more than 100%.

The results show that the default T_EX hyphenation rules perform quite well, with an overall accuracy of 92.5% across all word lists. The highest accuracy is observed in the ‘6 agrees’ category, with 98.1% correct hyphenations, indicating that the rules share a high correspondence with dictionaries on those words with strong consensus among them. However, there is a notable increase in incorrect and missing hyphenations as the consensus among sources decreases.

These findings highlight the robustness of the default rules in many cases but also reveal their limitations, particularly when dealing with less frequently hyphenated words. When regarding the frequency of occurrences of words in the Wikipedia corpus, the words with hyphenation errors amount to 3.7% of the occurrences in the corpus.

While the default T_EX hyphenation rules demonstrate a high level of accuracy, there remains room for improvement. The next sections will explore ways to enhance these rules, including the development of handcrafted patch rules and the use of Patgen to create a new set of rules from scratch. These efforts aim to further reduce hyphenation errors and improve the overall performance of the T_EX hyphenation system for Portuguese.

6 Handcrafted rules

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 T_EX hyphenation rules created by [8].

- 1 .g2no, .g2nó, .g2nô – *gnomo, gnóstico, gnômon*■
- 2 t2c – *tchau, tcheco*
- 3 1p2neu – *pneumonia, pneumotórax, pneumático, hidropneumático*
- 4 .t2m – *tmese*
- 5 .p2t – *ptose, pterossauro*
- 6 .m2n – *mnemônico*
- 7 c2za – *czar*
- 8 .s2 – *stalinismo*
- 9 .t2 – *tsunami, tzarista*
- 10 .p2si, .p2sí – *psicologia, psíquico*
- 11 su2b3r, su2b3l – *subrotina, sublunar*
exception: .su3b4li – *sublinhar, sublimar*
- 12 .ne4o – *neoliberal, neonazista*
- 13 1p2seuld – *pseudônimo*
- 14 1qu – *enquanto, inquieto, farquhar, qubit*
- 15 alir., ulir. – *sair, extrair, diminuir, incluir*
- 16 alind, ali1nh – *ainda, rainha*
- 17 elimp – *reimpresso, teleimpressor*
- 18 elinc, elinf, eling, elins, elint, elinv – *reincidência, reinfecção, reingressa, reinserção, reintegração, reinventar*
- 19 uliz., aliz. – *juiz, raiz*
- 20 proli1b – *proibição*
- 21 tu1i, bu1i, nu1i, olim, olin, ulin, su1i, í1e, ju1i, fu1i, du1i, dolim, auli, uli1ç – *intuitivo, contribuidor, ingenuidade, coimbra, coincide, ruindade, suicida, píer, juizado, fuinha, assiduidade, amendoim, cacaucultor, constituição*
exception: tu2id, tu2it, co2ima, o2i1na – *gratuidade, intuito, coima, boina*
- 22 al1ã, al1ã, al1é, al1í, al1ó, al1ô, al1ú, el1ã, el1ã, el1ã, el1é, el1ê, el1í, el1ó, el1ô, el1ú, é1o, í1ã, í1ã, í1é, í1í, í1ó, í1u, í1ú, í1a, í1o, ol1ã, ol1ã, ol1é, ol1ê, ol1í, ol1ó, ul1ã, ul1ã, ul1ã, ul1é, ul1ê, ul1í, ú1o – *abraâmico, abraão, aéreo, país, caótico, faraônico, saúde, balneário, oceânico, campeã*■
feérico, preênsil, veículo, teórico, napoleônico, conteúdo, néon, diário, região, soviético, údiche, periódico, feiura, viúva, maníaco, íon, razoável, joão, poético, boêmia, heroísmo, alcoólico, usuário■
itapuã, lituânia, suécia, cauê, suíça, flúor
exception: 1gu2ã, 1gu2ã, 1gu2é, 1gu2ê, 1gu2í, 1qu2ã, 1qu2ã, 1qu2ã, 1qu2é, 1qu2ê, 1qu2í, – *jaraguá, saguão, alguém, portugueses, linguística, aquático, camaquã, equânime, inquerito, sequência, química*
- 23 1bô, 1cô, 1çô, 1dô, 1fô, 1gô, 1lô, 1mô, 1nô, 1pô, 1rô, 1sô, 1tô, 1vô, 1xô, 1zô – *robô, recôncavo, maçônico, judô, telefônica, xangô, camelô, capô, sumô, econômico, tarô, subsônico, chatô, vovô, saxônia, amazônia*
- 24 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 notable that the latter examples are formed by the digraphs *qu-* and *gu-*, while the former involve vowels other than *i* and *u*. In Portuguese, a diphthong with an accented second vowel is only possible when preceded by *qu-* or *gu-*. The fourth group, in turn, which comprises rules 22 and 23⁴, which are counterparts of rules that were already in the default rules, but did not contemplate the cases with certain diacritics. 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	.p2sí	.g2no	.g2nó	.g2nô
t2c	1p2neu	.t2m	.p2t	su2b3r
su2b3l	.su3b4li	alir.	ulir.	1qu
lvô	llô	lcô	lgô	lbô
ltô	lrô	lpô	alé	alí
aló	alú	elá	elâ	elã
elé	elê	elí	eló	élo
elú	ilá	ilã	íla	ilé
ilí	iló	ílo	ilu	ilú
olá	olé	olí	oló	ulá
ulã	ulâ	ulí	úlo	1qu2ã
1qu2ã	1qu2í	1gu2í	alind	ali1nh
elimp	elinc	elinf	eling	elins
elint	elinv	uliz.	aliz.	4a.
4e.	4o.	1gu2á	1gu2ã	1qu2ã
.m2n	c2za	.s2	1p2seuld	ldô

⁴ Note that rule 22 is in the intersection between the third and fourth group of rules.

1fô	1mô	1nô	1sô	1zô
tu1i	tu2it	tu2id	buli	nuli
olin	ulin	suli	île	juli
fuli	du1i	do1im	auli	ulilç
u1ê	1gu2ê	1qu2ê	lçô	u1é
1gu2é	1qu2é	1xô	alâ	alã
alô	elô	.ne4o	olã	olê
olim	o2ilna	pro1ilb	co2ima	.t2

Rule 24 was created to inhibit final hyphenations, addressing errors arising from apparent proparoxytones. However, this resulted in missing hyphenations before a final vowel, which is a minor issue since such hyphenations are typically avoided to prevent orphan vowels⁵.

Tables 2 and 3 present the results for our patched hyphenation rules and for the rules with rule 24 removed, respectively.

Table 2: Results of patched T_EX hyphenation rules on the dictionaries created from online dictionaries hyphenations.

word list	correct	incorrect	missing	entries
6 agrees	96.6%	0.0%	3.4%	15,842
5 agrees	94.7%	0.1%	5.2%	15,642
4 agrees	94.1%	0.1%	5.9%	10,299
3 agrees	90.6%	0.5%	9.1%	7,766
total	94.6%	0.1%	5.4%	49,549

Some words might have a bad and a missing hyphenation point concomitantly, therefore the percentage in each line might add up to more than 100%.

Table 3: Results of patched T_EX hyphenation rules without three rules created to deal with apparent proparoxytones (rule 24).

word list	correct	incorrect	missing	entries
6 agrees	99.98%	0.0%	0.03%	15,842
5 agrees	92.7%	7.1%	0.3%	15,642
4 agrees	91.2%	8.2%	0.7%	10,299
3 agrees	98.2%	0.6%	1.4%	7,766
total	95.6%	4.0%	0.5%	49,549

Some words might have a bad and a missing hyphenation point concomitantly, therefore the percentage in each line might add up to more than 100%.

Considering the frequency of occurrences, the proportion of words with hyphenation errors has decreased from 3.7% to 2.5% in the Wikipedia corpus. While this represents a slight improvement, further reductions in errors may require a significant increase in the number and complexity of the rules, highlighting the diminishing returns on such refinements.

7 Patgen rules

Patgen was used to create new sets of rules from scratch, as well as rules starting from the default

⁵ Inhibit widows and orphans is also a T_EX’s directive.

ones or from the patched rules proposed in Section 6. This involves defining specific parameters, choosing a reference dictionary with correct hyphenations, and using a translation file tailored for Patgen.

There are infinite ways to specify the parameters for Patgen. Based on previous works, we opted to test two approaches: (1) using fixed parameters at all stages of Patgen execution (starting from an empty set or from the default rules) [10]; (2) using a progressive approach, starting from an empty set and creating higher-order rules at each step, always building on the rules of the previous step [31].

As a reference dictionary, we used the concatenation of the dictionaries with six, five, and four agreements. The last dictionary (three agreements) was used as a validation dictionary⁶.

For more information on how to use Patgen and create the translation file, we refer the reader to [10].

Table 4: Results of patched T_EX hyphenation rules created by Patgen on the validation dictionary with 7,745 entries.

model	correct	incorrect	missing	rules
fix [†] null	93.94%	1.26%	5.23%	849
fix [†] default	94.34%	0.84%	5.04%	1,091
fix [†] patched	94.46%	0.81%	4.94%	1,045
mfp [‡] null	8.36%	0.47%	91.56%	699
mfp [‡] default	94.23%	0.66%	5.32%	1,230
mfp [‡] patched	93.70%	1.00%	5.52%	1,230

Some words might have a bad and a missing hyphenation point concomitantly, therefore the percentage in each line might add up to more than 100%.

[†] fix null/default/patched refer to the method using fixed parameters with a given initial rules set.

[‡] mfp null/default/patched refer to the incremental approach using a given set of rules as a starting point.

Patgen creates a large set of rules that perform well within those specific sets, but have poor generalization capacity. Using smaller reference dictionaries will decrease the performance, but that is primarily a consequence of Patgen including long rules that resemble specific exceptions rather than general rules. Some examples include: **3linea**, **neári5** and **padri3**. It is also unrealistic to try to find a linguistic explanation for all rules created by Patgen, specially those involving large patterns.

Comparing the results from Table 4 with those from Tables 2 and 3, it becomes evident that Patgen’s performance falls short when evaluated using the same validation dictionary.

⁶ We have also tested to start with a smaller reference dictionary (using those with six and five agreements) and validate on the subsequent dictionaries, but the results are worse and therefore not discussed here.

8 Conclusion

Upon evaluating the effectiveness of the default \TeX hyphenation rules for Portuguese using our hyphenation dictionaries, several key insights emerged. The results, as shown in Table 1, highlight the performance metrics of the default rules and set the baseline for improvements. Our handcrafted rules, designed to address specific hyphenation errors, demonstrated superior performance and generalization capability, as shown in Tables 2 and 3. Patgen was utilized to generate new sets of rules, both from scratch and by building upon the default rules or our handcrafted rules. Despite the potential of Patgen, its performance was suboptimal when evaluated with the same validation dictionary, as observed in Table 4. This was primarily due to the creation of overly specific rules that failed to generalize well.

The comprehensive analysis reveals that while default and automatically generated rules have their merits, manually refined rules significantly enhance hyphenation accuracy. This enhancement is crucial for typesetting high-quality Portuguese texts, ensuring readability and maintaining aesthetic consistency. Although there remains potential for further refinement of the rules, it may not be worthwhile since our dataset already encompasses the majority of words in the Portuguese vocabulary (95% of all occurrences in Portuguese Wikipedia).

We believe that advanced hyphenation could require new algorithms or modifications to \TeX 's behavior, allowing it to account for other driving forces in hyphenation, such as: (1) Syllable-based hyphenation, (2) Compound words, (3) Morphological determination, (4) Context-sensitive hyphenation (5) Avoidance of stylistically undesirable hyphenations, (6) Prevention of ambiguities or unintended meanings, (7) Semantic-based hyphenation, and (8) Universal Syllabic Hyphenation, among others.

Future research should explore hybrid approaches that combine rule-based and machine learning techniques to further optimize hyphenation patterns. Additionally, expanding the dataset to include more diverse Portuguese corpora could help address rare edge cases and improve the overall robustness of the hyphenation system. We believe the findings of this study provide a solid foundation for ongoing improvements in \TeX hyphenation rules, ultimately contributing to better automated text processing in the Portuguese language.

References

- [1] Aulete. Aulete online dictionary. Accessed: 2023-06-26. <https://aulete.com.br/>

- [2] D.P. Cegalla. *Novíssima Gramática da Língua Portuguesa*. Companhia Editora Nacional, 2020.
- [3] C. Cunha, L. Cintra. *Nova gramática do português contemporâneo*. Lexikon Editora Digital, 2016.
- [4] P. da Língua Portuguesa. Portal da língua portuguesa online dictionary. Accessed: 2023-06-26. <http://www.portaldalinguaportuguesa.org/>
- [5] A.M. de Araújo, T. Maruyama. A hifenização em português. *Idioma*, (28):90–107, 2015. http://www.institutodeletras.uerj.br/idioma/numeros/28/Idioma28_a08.pdf
- [6] P. de Magalhães Gândavo. *Regras que ensinam a maneira de escrever e orthographia da lingua portuguesa, etc.* Lisboa, 1574. <https://purl.pt/324>
- [7] P.J. de Rezende. Portuguese hyphenation table for tex. *TUGboat* 8(2):102–102, 1987.
- [8] P.J. de Rezende, J.J.D. Almeida. Hyphenation patterns for portuguese. <http://mirror.ctan.org/language/hyph-utf8/tex/generic/hyph-utf8/patterns/tex/hyph-pt.tex>, 2015.
- [9] Dicio. Dicio online dictionary. Accessed: 2023-06-26. <https://www.dicio.com.br/>
- [10] Y. Haralambous. A revisited small tutorial on patgen, 28 years after. *electronic form, available from CTAN as info/patgen2. tutorial*, 2021.
- [11] Hunspell's Team. Hunspell. <http://hunspell.github.io/>
- [12] Hunspell's Team. Hunspell hyphen. <https://github.com/hunspell/hyphen>
- [13] D.E. Knuth. Preliminary preliminary description of TEX. Draft version, May 1977.
- [14] R. Levien. *Brief explanation of the hyphenation algorithm herein*. Hunspell, 1998. <https://github.com/hunspell/hyphen/blob/master/README.hyphen>
- [15] F. Liang, P. Breitenlohner. Pattern generation program for the tex82 hyphenator. Technical Report 2, Ctan, 1991. Electronic documentation of PATGEN.
- [16] F.M. Liang. *Word Hy-phen-a-tion by Computer*. Ph.D. thesis, Stanford University, 1983.
- [17] F.M. Liang. Word hy-phen-a-tion by computer, department of computer science, 1983.
- [18] Linguatca. Cetenfolha corpus. Accessed: 2023-06-26. <https://www.linguatca.pt/cetenfolha/>

- [19] Michaelis. Michaelis online dictionary. Accessed: 2023-06-26. <https://michaelis.uol.com.br/>
- [20] L. Németh. Automatic non-standard hyphenation in openoffice. org. *TUGboat* 27(1):32–37, 2006.
- [21] P. NET. Palavras net word list. Accessed: 2023-06-26. <https://www.palavras.net/>
- [22] T_EX pattern authors. T_EX hyphenation patterns. <https://www.tug.org/tex-hyphen/>
- [23] Priberam. Priberam online dictionary. Accessed: 2023-06-26. <https://dicionario.priberam.org/>
- [24] K. Scannell. Hyphenation patterns for minority languages. *TUGboat* 24(2):236–239, 2003.
- [25] Senado Federal. *Acordo ortográfico da língua portuguesa: atos internacionais e normas correlatas*. Senado Federal, Apr. 2013.
- [26] P. Sojka. Notes on compound word hyphenation in tex. *TUGboat* 16(3):290–297, 1995.
- [27] P. Sojka. Notes on compound word hyphenation in tex. Technical report, Masaryk University in Brno, Faculty of Informatics, August 1995.
- [28] P. Sojka. Hyphenation – a tutorial for TEX users. <https://www.fi.muni.cz/~sojka/PB029/hyptut.pdf>, 2002.
- [29] P. Sojka, D. Antoš. 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, 2003.
- [30] P. Sojka, et al. *Competing Patterns in Language Engineering and Computer Typesetting*. Ph.D. thesis, Masaryk University, Brno, 2005.
- [31] P. Sojka, O. Sojka. The unreasonable effectiveness of pattern generation. *TUGboat*, 40(2):187–193, 2019.
- [32] Wikcionário. Wikcionário online dictionary. Accessed: 2023-06-26. <https://pt.wiktionary.org>
- [33] Wikipedia. Portuguese wikipedia dump. Accessed: 2023-06-26. <https://dumps.wikimedia.org/ptwiki/latest/>

◇ Leonardo Araujo and Aline
 Benevides
 MG 443, Km 7 Fazenda do Cadete
 Ouro Branco, 36495-000
 Brazil
 leolca (at) ufsj dot edu dot br
<https://sites.google.com/site/leolca/>
 ORCID 0000-0003-3884-2177