Proposta recebida em Julho 2024 e aceite para publicação em Outubro 2024.

Aperfeiçoando a Hifenização Automática em Português no TEX

Enhancing Automatic Hyphenation in Portuguese for TEX

Leonardo Araujo
Universidade Federal de São João del-Rei

Aline Benevides
Faculdade de Tecnologia do Estado de São Paulo

Resumo

As regras de hifenização do português para o TEX têm sido utilizadas há mais de três décadas, apresentando um bom desempenho geral. Entretanto, ainda há hifenizações incorretas e pontos de hifenização não identificados. Esses pontos, embora em sua maioria ocorram perto das bordas das palavras e sejam irrelevantes para fins tipográficos no TEX, podem ser relevantes em contextos específicos, como ao lidar com palavras fora do léxico padrão ou com aplicações que fazem o uso da segmentação silábica/tipográfica. A partir de uma análise de 49.528 palavras hifenizadas, obtidas de dicionários online, propusemos 120 novas regras para serem incorporadas às regras existentes de hifenização do português. Além disso, utilizamos o *patgen* para criar novas regras ou melhorar as já existentes. No entanto, as regras geradas pelo *patgen* não demonstraram boa capacidade de generalização. Em última análise, as regras manuais ajustadas apresentaram o melhor desempenho, resultando em um aumento de 2.1% na taxa de acertos. O número de pontos de hifenização corretos aumentou de 38.519 para 39.808, enquanto os pontos de hifenização incorretos diminuíram drasticamente de 2.059 para 33. Importante ressaltar também que as regras elaboradas manualmente demonstraram uma melhor capacidade de generalização do que as regras geradas automaticamente pelo *patgen*.

Palavras chave

hifenização; padrões de hifenização; hifenização automática em Português

Abstract

Portuguese hyphenation rules for TEX have been in use for over three decades, showing good overall performance. However, there are still incorrect hyphenations and undetected hyphenation points. These points, although mostly occurring near word boundaries and being irrelevant for typographic purposes in TEX, can be relevant in specific contexts, such as when dealing with words outside the standard lexicon or in applications that utilize syllabic/typographic segmentation. Based on an analysis of 49,528 hyphenated words obtained from online dictionaries, we proposed 120 new rules to be incorporated into the existing Portuguese hyphenation rules. Additionally, we used *patgen* to create new rules or improve existing ones. However, the rules generated by *patgen* did not demonstrate good generalization capability. Ultimately, the manually adjusted rules showed the best performance, resulting in a 2.1% increase in the success rate. The number of correct hyphenation points increased from 38,519 to 39,808, while the incorrect hyphenation points drastically decreased from 2,059 to 33. It is also important to note that the manually crafted rules demonstrated better generalization capability than the automatically generated rules by *patgen*.

Keywords

hyphenation; hyphenation patterns; automatic hyphenation in Portuguese

1. Introdução

Durante muito tempo, o hífen não era utilizado na quebra de linha, de forma que as palavras precisavam caber inteiramente em uma linha ou eram cortadas em locais arbitrários. Inicialmente, nenhuma marcação era utilizada para indicar a quebra de linha, o que poderia causar confusão ou alguma interpretação indesejada. Como resultado, os ortógrafos advogaram pela introdução de um sinal para indicar essas quebras. O português enfrentou a mesma introdução gradual de um sinal de hifenização para marcar palavras que se estendiam por linhas. Embora o uso de um sinal de hifenização (=) tenha sido defendido pelos ortógrafos (Gândavo, 1981), poucos documentos utilizavam esse sinal até o final do século XVIII (de Araújo & Maruyama, 2015).

A hifenização pode dificultar a leitura fluida e, por isso, em alguns casos deve ser evitada. Na literatura infantil, por exemplo, pode interromper o fluxo e a compreensão da leitura. Da mesma forma, em fontes grandes, títulos ou manchetes, a hifenização parece visualmente desagradável. Em documentos técnicos, pode causar confusão quando aplicada a jargões. Por outro lado, espaços grandes ou pequenos entre pala-

vras também podem dificultar o processo de leitura, tornando a hifenização fundamental quando os textos usam comprimentos de linha curtos. À medida que uma linha fica mais curta, o número de candidatos à quebra entre as palavras diminui, levando a espaços irregulares entre as palavras e os caracteres. Em textos alinhados à esquerda, o espaçamento consistente pode eliminar o risco de formação de rios (espaços brancos irregulares), garantindo uma aparência consistente e esteticamente agradável do texto. Da mesma forma, em texto justificado, a hifenização pode criar espaçamento irregular, reduzindo a legibilidade. Portanto, a hifenização automática desempenha um papel importante na tipografia.

O TEX, como sistema de composição tipográfica, aborda meticulosamente essas questões. Através da hifenização automática e do ajuste fino do espaçamento entre pares específicos de letras (chamado de *kerning*), ele organiza o texto na página para proporcionar uma leitura fluida e visualmente agradável. Esse processo considera diversos fatores que influenciam a legibilidade, como altura e comprimento das linhas, tamanho dos parágrafos, tipo e tamanho da fonte, e espaçamento entre letras e palavras.

O espaçamento ideal entre palavras evita a formação de 'lagos' (espaços excessivos) e 'rios' (colunas irregulares de espaços brancos), garantindo clareza e conforto visual. Além disso, o TEX previne 'órfãos' (última linha de um parágrafo com apenas uma ou poucas palavras) e 'viúvas' (primeira linha de um parágrafo sozinha na página anterior), melhorando o fluxo do texto. O controle automático dos espaços em branco entre caracteres, palavras, linhas, parágrafos, seções e recuos proporciona um espaçamento ideal entre os elementos do texto, garantindo uma distribuição uniforme na página, evitando problemas de legibilidade e estética.

O presente trabalho se propõe a analisar as regras de hifenização automática do TEX, revisitar a utilização do programa *patgen* para criar novas regras, e analisar as regras de hifenização do português, comparando aquelas propostas por de Rezende (1987); de Rezende & Almeida (2015) e as atualizações propostas por Araujo & Benevides (2024). Para realizar tal comparação, será utilizado como referência um dicionário de hifenização do português criado a partir de seis recursos lexicais online Michaelis (sd.), Priberam (sd.), Wikcionário (sd.), Aulete (sd.), Portal da Língua Portuguesa (sd.) e Dicio (sd.).

Araujo & Benevides (2024) propõem a atualização das regras de hifenização para o TEX, utilizando regras elaboradas manualmente e comparando-as com aquelas geradas pelo patgen, destacando a limitada capacidade de generalização deste último. O presente artigo aprofunda e amplia o escopo da análise anterior, abordando o tema sob uma perspectiva mais abrangente da língua portuguesa, com foco nos aspectos fonológicos, gramaticais e estilísticos. Além de esclarecer as escolhas adotadas, o estudo aponta direções para futuras melhorias e discute as limitações inerentes ao processo de hifenização automática.

2. Hifenização

Os idiomas, em geral, podem ter suas hifenizações guiadas de quatro critérios distintos ou pela relação entre elas: (1) fonológico, baseada na divisão silábica da fala; (2) morfológico (ou etimológico), focada nas partes que proveem significado às palavras (prefixos, radicais e sufixos); (3) ortográfico, seguindo convenções de escrita padrão; e (4) semântico, que considera o contexto para evitar cortes ambíguos ou inadequados.

A regra geral de hifenização no português é dividir nas fronteiras entre sílabas. Uma sílaba é formada por um núcleo obrigatório, que, no português, pode ser preenchido apenas por vogal, e consoantes periféricas opcionais (antes ou depois do núcleo). Em algumas situações, a divisão silábica não respeita os constituintes morfológicos, criando uma relação conflitante com o padrão morfológico, que também desempenha papel fundamental na hifenização no português. Prefixos como *bis-*, *des-* e *in-* são exemplos: de um lado, temos hifenizações, como *bis-ne-to*, *des-co-brir* e *in-va-ri-á-vel*, em que as fronteiras morfológicas são respeitas; de outro lado, temos *bi-sa-vô*, *de-sem-bar-que*, *i-na-ti-vo* e *i-nobs-tan-te*[1], em que os prefixos são divididos em duas sílabas—neste caso, a hierarquia de sonoridade dita a preferência por preencher o ataque sílaba, e não a coda.

Em alguns casos, a hifenização também é uma questão de estilo, tendo em vista que algumas divisões soam melhores que outras. Essas alternativas conflitantes geralmente surgem quando uma palavra possui muitos pontos de hifenização possíveis. Considere *ar-que-ó-lo-go*, que é preferencialmente particionado como *arque-ólogo* em oposição a *arqueó-logo*. É preferível manter morfemas inteiros juntos: *arque* (que significa an-

[1] A regra de divisão silábica pode levar a duas partições possíveis, baseadas em forças fonológicas ou morfológicas: *i-nobs-tan-te* e *in-obs-tan-te*, mas a primeira é preferível.

tigo, primitivo) e *ólogo* (aquele que estuda o assunto). Em português, a palavra *en-tres-sa-fra* é preferencialmente separada como *entres-safra* em oposição a *en-tressafra* ou *entressa-fra*, e a palavra *fe-liz-men-te* é preferencialmente separada como *feliz-mente* em oposição a *fe-lizmente* ou *felizmen-te*. Em ambos os casos, manter os dois morfemas juntos foi a principal motivação. Também é mais elegante evitar divisões entre consoantes ou vogais duplas, mesmo que exista um ponto de hifenização entre essas letras. Por exemplo, *pressu-rizar* é preferível a *pres-surizar* e *empreen-dedor* é preferível a *empre-endedor*. Ainda assim, existem exceções, sendo preferível particionar *micro-organismo* em vez de *microorganismo* (manter morfemas juntos é preferível a dividir uma vogal dupla).

A hifenização ortográfica considera como as palavras são escritas e segue as convenções de um determinado idioma. Em português, a hifenização evita separar dígrafos quando formados por consoantes distintas, como *lh* em *mi-lho* e *nh* em *ba-nho*; mas as separam quando constituídas por consoantes iguais, como *ss* em *as-sim* e *rr* em *ter-ra*. Também tende a manter certos agrupamentos consonantais juntos, como *br* em *cobra* (*co-bra*) e *tr* em *atração* (*a-tra-ção*). Há, ainda, casos como *bíceps* (*bí-ceps*) e *advogado* (*ad-vo-ga-do*), com consoantes mudas na grafia, mas com epêntese na fala. As sequências consonantais constituiriam sílabas próprias, entretanto a ortografia impede que sejam separadas deixando uma sequência não permitida pela fonotática da língua (*bí-ce-ps* e *a-d-vo-ga-do*, respectivamente).

Uma palavra com diferentes significados pode ser hifenizada de formas distintas de acordo com o sentido. Por exemplo, a palavra sueca *glassko* possui três significados e pode ser hifenizada como *glas-sko* (sapato de vidro), *glass-ko* (vaca de sorvete) e, de forma não padrão, *glass-sko* (sapato de sorvete) (Németh, 2006). Em português, a palavra *sublinha* pode ser hifenizada de duas maneiras: *su-bli-nha* (verbo) quando representa a forma flexionada do verbo *sublinhar* ou *sub-linha* (substantivo) quando se refere ao traço sublinhado.

Outro ponto importante a considerar são as ambiguidades que podem surgir quando uma palavra é particionada durante o processo de hifenização. No âmbito da composição tipográfica, em português, devemos evitar hifenizações como *de-putada*, *fede-ração*, *acu-mula*, *após-tolo* e *cú-bico*. Hifenizações que possam levar o leitor a pronunciar incorretamente uma palavra também devem ser evitadas, como *pe-rigo*.

Diante disso, a automatização desse processo pode se dar de duas maneiras: i. a partir de uma abordagem baseada em dicionário, que restringe as possibilidades de hifenização às entradas presentes em um dicionário; ou ii. a partir de uma abordagem baseada em algoritmo, que pode ser aplicada a qualquer sequência encontrada no texto. Um modelo baseado em regras pode incluir o reconhecimento de prefixos, sufixos, morfemas ou sequências adequadas para a inserção de um hífen. A abordagem baseada em algoritmo pode utilizar um sistema lógico para analisar palavras e aplicar as regras de hifenização de um determinado idioma. Como as regras de hifenização variam significativamente entre línguas, é necessário desenvolver um algoritmo para cada uma. Embora um sistema baseado em lógica possa ser eficiente e compacto, ele ainda precisa lidar com exceções por meio de regras pré-programadas. Outra abordagem é o uso de padrões de correspondência. Ao utilizar um corpus com exemplos de hifenização em um idioma, essa abordagem identifica sequências de letras que determinam pontos de hifenização adequados. Os padrões podem englobar prefixos, sufixos, exceções e regras especiais de hifenização do idioma. Por fim, um modelo híbrido também pode ser utilizado. Ele combina dois ou mais dos métodos descritos anteriormente para aprimorar a precisão e flexibilidade da hifenização.

3. Hifenização automática no TEX

A hifenização automática é um componente crucial dos sistemas de preparação de documentos, especialmente na composição tipográfica e na editoração eletrônica. Ela contribui para a legibilidade e estética do texto. Ao automatizar a inserção de hifens em locais adequados, dispensa-se a intervenção de editores ou tipógrafos e, ainda, pode ser facilmente adaptada para diferentes idiomas.

Apesar do fato de o algoritmo e as regras de hifenização do TEX serem antigos, eles são, até hoje, a abordagem mais usada, mesmo fora do mundo do TEX. A base para isso é o Hunspell, um verificador ortográfico e analisador morfológico adotado por diversos softwares (por exemplo, LibreOffice, Mozilla Firefox, Mozilla Thunderbird, Google Chrome, macOS, InDesign, memoQ, Opera, Affinity Publisher, entre outros (Hunspell's Team, 2023)). O Hunspell usa as regras de hifenização do TEX (Hunspell's Team, sd.; Levien, 1998), tornando a hifenização do TEX amplamente difundida no mundo da computação. Isso é resultado da simplicidade e versatilidade da abordagem do TEX. O algoritmo funciona de maneira eficaz, pois já suporta re-

gras para 66 idiomas (TEX pattern authors, sd.) e oferece flexibilidade para criar regras para qualquer idioma atualmente não suportado.

A hifenização automática no TEX foi introduzida por Knuth (1977). O seu desenvolvimento focava nas regras para o inglês, as quais compreendiam: (1) cada parte deve conter uma vogal, exceto *e* final; (2) remoção de sufixo; (3) remoção de prefixo; e (4) quebra vogal-consoante-consoante-vogal (VCCV) (mas combinando *h* com a letra anterior se for consoante). Testes mostraram que este primeiro algoritmo poderia encontrar 40% dos pontos de hifenização permitidos (Liang, 1983). O algoritmo de hifenização proposto por Liang (1983) e adotado no TEX usa a noção de padrões concorrentes (Liang, 1983). Um banco de dados de palavras hifenizadas é varrido em busca de padrões de hifenização e inibição. O algoritmo introduzido no TEX82 utiliza cinco níveis alternados de padrões de hifenização e inibição. O *patgen*, programa para geração de padrões baseado em um corpus, foi criado por Liang (1983) (Liang & Breitenlohner, 1991) e usado para criar padrões de hifenização para vários idiomas (Sojka, 1995a,b; Sojka et al., 2005; Sojka & Antoš, 2003; Scannell, 2003).

A hifenização efetiva de palavras pelo TEX depende dos seguintes fatores: (1) idioma do documento, que determina qual conjunto de padrões aplicar; (2) caracteres usados, pois alguns podem bloquear a hifenização em suas extremidades; (3) o valor das variáveis internas `\lefthyphenmin` e `\righthyphenmin`, que definem o comprimento mínimo da sequência de caracteres nas bordas esquerda e direita antes que qualquer hifenização seja permitida (Sojka, 2002).

Para simplificar, se considerarmos apenas o alfabeto latino, sem caracteres diacríticos, os padrões usados na hifenização do TEX têm a forma: `^\.?[0-9]?([a-z]+[0-9]?)+\.?\$`, sendo esta uma expressão regular.Um exemplo desse padrão é `4z1z2` (parte das regras de hifenização do inglês), que é composto por uma sequência de letras e números. Em geral, usamos caracteres/símbolos do alfabeto da língua, junto com números indo-arábicos, para expressar facilitação ou inibição de hifenização. Números ímpares indicam um bom ponto de hifenização, enquanto números pares indicam um local inadequado para quebra. O exemplo dado indica que a sequência tem um bom ponto de quebra entre o primeiro e o segundo *z* e que a hifenização deve ser inibida antes do primeiro *z* e após o segundo *z*. Por exemplo, a hifenização das palavras *piz-za*, *fiz-zle* e *mezzanine* segue essa regra, em que o hífen é colocado entre os dois *z* e nenhum hífen antes nem depois dos *z*. Os padrões também podem usar o ponto (.) para indicar limites de palavras. O padrão `.sh2` se aplica ao início das palavras, o que implica que o *s* e o *h* devem ficar juntos no início de uma palavra e a hifenização também deve ser inibida após o *h*. Por exemplo, este padrão é utilizado em *Sher-lock*.

As regras de hifenização são organizadas em níveis, de 1 a 9, onde números ímpares representam níveis de hifenização e números pares representam níveis de inibição. Cada nível funciona como um nível de exceção para o seu predecessor. Por exemplo, a regra `sh1er` indica um bom ponto de hifenização entre o *h* e o *e* na sequência *sher*. Uma regra em um nível superior, como `.sh2`, implica uma exceção à regra do nível inferior. Quando vemos *sher* no início de uma palavra, a regra `.sh2` se aplica e a hifenização proposta pela regra do nível inferior `sh1er` deve ser impedida. Esse é o caso da hifenização da palavra *Sher-lock*. O exemplo completo é fornecido na Listagem 1[2], na qual é possível verificar todas as regras pertinentes do inglês atuando na hifenização de *Sher-lock*.

```
.  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  -
```

Listagem 1: Exemplo das regras aplicadas à hifenização da palavra *Sherlock*.

Algumas regras de hifenização não podem ser implementadas usando o algoritmo de hifenização do TEX, tendo em vista que, no alemão, por exemplo, a hifenização pode levar à alteração ou à inserção de letras. Além disso, palavras compostas não possuem hifens, resultando em sequências longas de letras sem separação visível e até mesmo repetições da mesma letra, como visto em *Wasserrinne* e *Schifffahrt*. A reforma ortográfica do alemão também introduziu algumas mudanças, tornando necessária a criação de um conjunto diferente de regras de hifenização. Por exemplo, a palavra *Schiffahrt* deve ser hifenizada como *Schiff-fahrt*, preservando os *f*s de cada palavra que constituem esse composto. A hifenização deve inserir um *f* que não faz parte da forma escrita, o que não era um problema para a antiga forma escrita da palavra: *Schifffahrt*. Além disso, as antigas regras de hife-

[2]Este exemplo foi criado utilizando-se o algoritmo de hifenização do TEX portado para a linguagem Go. Seu código está disponível em https://github.com/speedata/hyphenation.

nização da gramática alemã estabeleciam a hifenização *Bäk-ker* para a palavra *Bäcker*, *Zuk-ker* para a palavra *Zucker* e *pak-ken* para a palavra *packen*. Atualmente, essas palavras são hifenizadas como *Bä-cker*, *Zu-cker* e *pa-cken*, respectivamente. Algumas palavras também mudaram seu ponto de hifenização após a reforma ortográfica: por exemplo, *Fen-ster* tornou-se *Fens-ter* e *meistens* tornou-se *meis-tens*. Alguns problemas na hifenização de palavras compostas no TEX são discutidos em Sojka (1995b).

Em síntese, um padrão consiste em uma sequência de caracteres (do alfabeto da linguagem) podendo conter um número intercalado, expressando o nível de hifenização/inibição. O marcador de limite de palavra, isto é, o ponto final, pode ser usado ocasionalmente nas extremidades do padrão. Quando não há número entre os caracteres em um padrão, assume-se o valor zero, o que significa *indefinido* e nenhum ponto de hifenização será sugerido nessa localização.

3.1. Patgen

Patgen utiliza uma lista de palavras hifenizadas para extrair padrões e, a partir deles, definir regras em vários níveis e com diferentes comprimentos. Ele começa com padrões curtos e aumenta incrementalmente seu comprimento até atingir o comprimento máximo permitido pelo usuário. O objetivo é manter os padrões o mais concisos possível, para obter uma melhor generalização. À medida que avança e padrões mais longos são incorporados, o *patgen* estabelece exceções. Em certos casos, pode ser necessária a análise de padrões longos, pois alguns pontos de hifenização podem depender de caracteres distantes do ponto de quebra.[3]

O *Patgen* trabalha com índices de *glifo*[4] em vez de códigos de caracteres. Cada glifo é representado por um único byte. Isso resulta em 256 índices, onde 13 deles são reservados para os dígitos 0-9 e os caracteres '.', '-' e '*'. Os 243 restantes são usados para representar símbolos de uma determinada língua. Os dígitos 0-9 são reservados para expressar níveis de regras de hifenização, e os caracteres '.', '-' e '*' são reservados para expressar um ponto de hifenização incorreto, um ponto de hifenização ausente e um ponto de hifenização correto, respectivamente. Para executar o *patgen*, é necessário um arquivo de tradução. Este arquivo define os valores de certos parâmetros específicos do idioma (na primeira linha) e enumera as várias formas como os símbolos do idioma podem aparecer (todas as linhas subsequentes). Na primeira linha, as posições 1 e 2 são usadas para definir o valor de `lefthyphenmin`, e as posições 3 e 4 são usadas para definir o valor de `righthyphenmin`. Esses valores determinam o comprimento mínimo de uma sequência de caracteres que pode ser gerada por um procedimento de hifenização. Para definir um valor de um único dígito, deixe a primeira posição em branco, ou seja, coloque um espaço nas posições 1 e 3 para `lefthyphenmin` e `righthyphenmin`, respectivamente. As posições 5, 6 e 7 são usadas para definir valores alternativos para os caracteres especiais '.', '-' e '*'. Um esquemático da primeira linha do arquivo de tradução é apresento na Figura 1.

As linhas a seguir usam um delimitador para demarcar cada letra do alfabeto da linguagem desejada, incluindo suas representações alternativas. A primeira posição da linha define o delimitador, e cada símbolo da linguagem pode ocupar tantas posições quanto necessárias, desde que o valor reservado para o delimitador não seja usado na definição do símbolo. As linhas de definição terminam quando o delimitador aparece duas vezes consecutivas. Considere o seguinte exemplo para definir a letra *e*, fazendo-se a suposição de equivalência entre as várias formas em que tal letra pode aparecer (minúsculas, maiúsculas, com ou sem diacríticos):

```
XeXEXéXêXÉXÊX\'{e}X\^{e}X\'{E}X\^{E}XX.
```

Este exemplo está ilustrado na Figura 2 para uma melhor compreensão. Adotamos `X` como delimitador. Observe que usamos a entrada direta (usando UTF-8 ou outra codificação que suporte o caractere *e* acentuado) e também a contraparte composicional usando a sequência de controle TEX apropriada que instrui o TEX a colocar diacrítico no caractere. Nesta definição, assumimos que as várias formas nas quais podemos encontrar o caractere *e* serão equivalentes para fins de hifenização (padrão de correspondência). Vale ressaltar que não é este o caso do português, o exemplo tem caráter meramente ilustrativo. O acento, muitas vezes representado graficamente pela presença de diacrítico, possui papel preponderante na determinação de sílabas e, consequentemente, na hifenização no português. Como outro exemplo, veja a linha que define o caractere π (Haralambous, 2021): `#p#P#\varpi ##.`

[3] Alguns exemplos, em inglês, de dependência de hifenização em caracteres distantes do ponto de quebra são: *dem-o-crat* e *de-moc-ra-cy*; *as-pi-rin* e *aspir-ing*; *de-monstra-tive* e *dem-on-stra-tion*.

[4] O termo *glifo* é usado comumente em linguística, tipografia e computação gráfica para se referir a uma representação gráfica específica de um caractere ou símbolo, que pode ser o símbolo inteiro ou um elemento visual distinto dentro dele.

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

Figura 1: Primeira linha do arquivo de tradução utilizado pelo *patgen*.

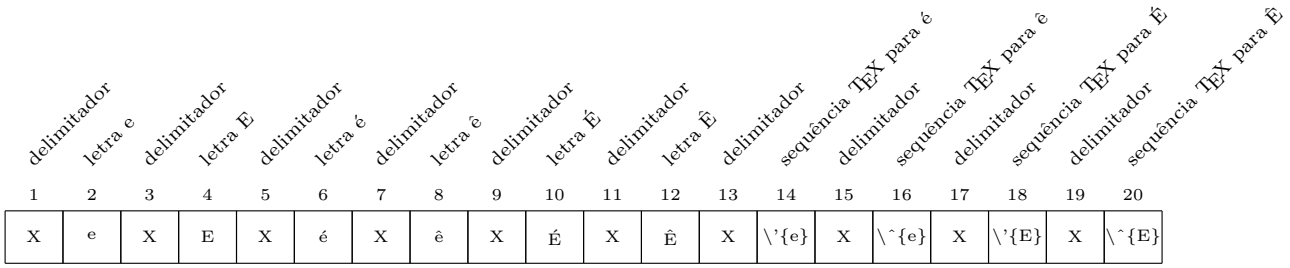

Figura 2: Exemplo de uma linha do arquivo de tradução onde são definidas as várias formas equivalentes para a letra *e*.

O *Patgen* também necessita de um arquivo de dicionário, que é uma lista de palavras pré-hifenizadas da qual o *Patgen* extrai padrões para criar regras de hifenização. Para garantir a funcionalidade adequada do *Patgen*, o arquivo de tradução e o arquivo de dicionário devem utilizar a mesma codificação, mesmo que seja uma codificação de vários bytes. O arquivo de tradução descreve como lidar com sequências de bytes que representam um glifo, e o *Patgen* funcionará perfeitamente quando houver no máximo 243 símbolos no idioma especificado.

A sintaxe para executar o *patgen* é apresentada na Listagem 2.

```
patgen arquivo_dicionario
   padroes_iniciais arquivo_saida
   arquivo_traducao
```

Listagem 2: Sintaxe para executar o *patgen*.

O arquivo de dicionário é uma lista de palavras hifenizadas corretamente, uma por linha; o arquivo de padrões iniciais é um conjunto de regras de hifenização a ser usado como ponto de partida; o arquivo de saída é o conjunto de regras finais criadas pelo *patgen*; e o arquivo de tradução, que mapeia as várias formas como cada símbolo de um idioma pode aparecer em documentos TEX. Se desejar executar o *patgen* do zero, começando com um conjunto vazio de regras, basta usar um arquivo vazio como o conjunto de padrões iniciais.

O *Patgen* utiliza alguns parâmetros ao longo de sua execução:

hyph_start, hyph_finish Dois números entre 1 e 9 (separados por um espaço) representam os níveis de padrão desejados no conjunto final de regras. Níveis de padrão ímpares são níveis de hifenização e níveis de padrão pares são níveis de inibição. Valores de nível mais alto prevalecem sobre os mais baixos, criando exceções, exceções sobre exceções e assim por diante. `hyph_start` e `hyph_finish` especificam o primeiro e o último nível, respectivamente, a serem considerados durante o processo de criação de regras.

pat_start, pat_finish Os padrões em cada nível são escolhidos em ordem crescente de tamanho (geralmente começando com tamanho 2). Isso é controlado pelos parâmetros `pat_start` e `pat_finish` especificados no início de cada nível. Estes representam o comprimento mínimo e máximo dos padrões de nosso interesse. Seus valores variam entre 1 e 15.

good weight, bad weight, threshold Cada nível de padrões é testado em todas as palavras do dicionário. Um padrão só é aceito se satisfaz a seguinte condição: $\alpha \times \#\text{good matches} - \beta \times \#\text{bad matches} \geq \eta$, onde α é o `good weight` (peso bom), β é o `bad weight` (peso ruim) e η é o `threshold` (limiar).

4. Sistemas de escrita

Existem diferentes sistemas de escrita, categorizados em logográficos, silábicos e alfabéticos. Embora distintos, esses sistemas podem ser construídos a partir da interação entre essas categorias (Coulmas, 2003; Palmer, 2010). Os princípios que orientam um sistema de escrita alfabético ou fonêmico variam significativamente de acordo com a língua e sua história. Aqui estão os principais pontos a serem destacados:

Representação fonêmica O princípio mais comum envolve representar os sons de uma língua por símbolos escritos, usando um

símbolo ou uma combinação de símbolos para representar cada som (por exemplo, sistema ortográfico do português usa um alfabeto de 26 letras e símbolos diacríticos que podem modificar o som de algumas letras ou indicam diferenças de pronúncia e acentuação, como o acento agudo (´), o acento grave[5] (`), o acento circunflexo (ˆ), o til (˜) e a cedilha (ç)).

Etimologia A etimologia também desempenha um papel crucial, já que a grafia de uma palavra reflete sua origem e seu desenvolvimento histórico (por exemplo, a língua francesa muitas vezes reflete as raízes latinas das palavras em sua grafia).

Morfologia A morfologia serve como guia para estruturar muitas línguas, usando letras ou símbolos específicos para indicar terminações, prefixos ou sufixos de palavras (por exemplo, o russo usa diferentes formas do alfabeto cirílico para indicar gênero e caso em seus substantivos).

Evolução da língua A evolução de uma língua ao longo do tempo também contribui para sua forma escrita (por exemplo, a grafia de muitas palavras em inglês mudou ao longo do tempo para refletir as mudanças na pronúncia).

Os sistemas ortográficos de línguas com escrita logográfica, como ideogramas chineses, escrita cuneiforme e hieróglifos egípcios, diferem significativamente dos sistemas de escrita alfabética. Nos sistemas logográficos, símbolos ou caracteres individuais representam palavras ou ideias inteiras, e não fonemas ou sons. Como resultado, os princípios que orientam seus sistemas ortográficos baseiam-se mais em conceitos semânticos e visuais do que em princípios fonêmicos. A hifenização é uma característica das escritas alfabéticas, uma vez que nesses sistemas palavras são formadas a partir de letras. Espaços ou hifens são usados para separar palavras, sílabas ou partes de palavras. Sistemas de escrita logográfica, como o chinês, funcionam de maneira diferente e não utilizam hifenização (Honorof & Feldman, 2006).

4.1. Sistema de escrita do português

O português utiliza um sistema de escrita alfabética, o que significa que sua ortografia é orientada por princípios fonológicos (Cagliari, 2015). A correlação entre escrita e pronúncia influencia a hifenização das palavras, pois elas são divididas em sílabas com base no sistema fonêmico. É importante observar que diferentes idiomas seguem princípios diversos para a divisão de palavras. Por exemplo, o inglês é guiado principalmente por princípios morfológicos, evidentes em palavras como *walk-ing*, *un-happy*, *work-s* e *ear-ly*. Além disso, outros fatores também influenciam a hifenização em inglês, como a distinção entre vogais longas e curtas, que funcionam no contexto de sílabas abertas ou fechadas, respectivamente; e a presença de consoantes duplas e dígrafos (Lin, 2011; Yavas, 2020). Embora cada idioma tenha seus princípios orientadores para o processo de hifenização, múltiplos fatores entram em jogo, levando a várias soluções em certos cenários. Por exemplo, em português, o princípio fonológico levaria a *hipe-rativo*, enquanto o princípio morfológico levaria a *hiper-inflação*. Ambas as abordagens parecem válidas e expressam comportamentos linguísticos distintos, em decorrência de questões semânticas[6]. Algumas palavras raras, como *hiperalgesia*, podem não estar sujeitas a influências morfológicas. Como tem baixa frequência de ocorrência, o indivíduo pode não estar ciente de seus componentes morfológicos e, portanto, hifenizá-la como *hipe-ralgesia*. A forma *hiper-algesia* também poderia ser aceita, enfatizando seus constituintes morfológicos. Além disso, existem inúmeras exceções que podem ser categorizadas em regras específicas.

O simples fato de um sistema ortográfico ser guiado por questões fonológicas não significa necessariamente que suas regras de hifenização espelhem diretamente a contraparte fonética. Isso é particularmente evidente no português, já que não há uma correspondência estrita entre letras e sons. O sistema ortográfico opera de acordo com suas regras próprias. Por exemplo, considere grupos consonantais que criam um único som (dígrafos) em palavras como *a<u>ch</u>ado*, *i<u>lh</u>a*, *su<u>sh</u>i*, *ca<u>rr</u>o* e *ma<u>ss</u>a*. Embora esses dígrafos sejam pronunciados como um único som dentro de uma única sílaba, sua representação na escrita determina como são divididos. Especificamente, consoantes diferentes dentro de um dígrafo devem permanecer juntas, enquanto consoantes idênticas são separadas. Como resultado, observamos hifenizações como *a-cha-do*, *i-lha* e *su-shi*, mas *car-ro* e *mas-sa*.

[5] O acento grave no português não modifica sons. Ele é utilizado para marcar a contração de duas vogais iguais.

[6] Observando palavras com prefixos, parece que mesmo nessas palavras a abordagem fonológica foi predominante na hifenização, mas não podemos dizer se é um subproduto da hifenização automática que pode ser usada para hifenizar palavras em dicionários online.

Em português, o hífen é permitido nos limites das sílabas e, em geral, segue princípios fonológicos. No entanto, de acordo com as gramáticas (Cunha & Cintra, 2016; Bergström & Reis, 2011; Cegalla, 2020), algumas regras específicas ainda podem se aplicar:

Regras de Não-Separação

1. ditongos e tritongos não devem ser separados (por exemplo, *mui-to*, *Pa-ra-guai*);
2. as sequências *ia*, *ie*, *io*, *oa*, *ua*, *ue* e *uo*, quando em posição final átona, não devem ser separadas (por exemplo, *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. grupos consonantais iniciais de sílaba não devem ser separados (por exemplo, *pneu-má-ti-co*, *psi-có-lo-go*, *mne-mô-ni-co*);
4. os dígrafos *ch*, *lh*, *nh* não devem ser separados (por exemplo, *ra-char*, *a-bro-lhos*, *ma-nhã*);
5. bigramas como *gu* e *qu*, cujo *u* não tem valor fonético, nunca são separados da vogal ou ditongo que os seguem (por exemplo, *U-ru-guai*, *pe-que*);
6. como formam um dígrafo, a vogal e a marca de nasalização subsequente (*m* ou *n*)não devem ser separados (por exemplo, *am-bição*, *man-cha*);
7. ditongos decrescentes não devem ser separados (por exemplo, *ai-ro-so*, *cau-te-la*, *ca-dei-ra*, *cha-péu*, *o-ra-ção*, *noi-te*, *ca-la-bou-ço*, *as-te-rói-de*, *re-tri-bui*);
8. ditongos crescentes não devem ser separados (por exemplo, *his-tó-ria*, *má-goa*, *sé-rio*, *gló-ria*, *fre-quen-te*, *pá-tria*);
9. dissílabos com apenas uma vogal por sílaba não devem ser quebrados (por exemplo, *ato*, *rua*, *ódio*, *unha*);
10. palavras com mais de duas sílabas, ao serem separadas, não podem deixar uma sílaba formada por uma vogal sozinha (por exemplo, *agos-to*, *la-goa*, *ida-de*);

Regras de Separação

11. vogais em hiato e sequências vocálicas em que cada vogal pertence a uma sílaba diferente devem ser separadas (por exemplo, *sa-ú-de*, *ra-i-nha*, *do-er*, *vo-os*), o mesmo procedimento é usado para separar ditongos em sílabas diferentes (por exemplo, *cai-ais*), ou ditongo e vogal em sílabas diferentes (por exemplo, *en-sai-os*);
12. sequências consonantais, quando em sílabas diferentes, devem ser separadas (por exemplo, *af-ta*, *ab-di-car*, *res-ci-são*, *ab-so-lu-to*);
13. os seguintes dígrafos consonantais devem ser separados: *rr*, *ss*, *mm*, *nn*, *sc*, *sç* e *xc* (por exemplo, *ter-ra*, *pro-fes-sor*, *co-mum-men-te*, *con-nos-co*[7], *des-cer*, *cres-ça*, *ex-ce-der*).

Regras 9 e 10 têm como objetivo principal garantir a legibilidade adequada do texto, alinhando-se com a abordagem do TEX para lidar com viúvas e órfãs tipográficas. Como mencionado na Seção 3, as variáveis `lefthyphenmin` e `righthyphenmin` controlam o tamanho mínimo para fragmentos de palavras hifenizadas. Quando essas variáveis são definidas com valores maiores que um, as regras 9 e 10 se tornam regras desnecessárias na hifenização do TEX. No entanto, a hifenização completa das palavras pode ser útil, principalmente em aplicativos de conversão de texto em voz (Libossek & Schiel, 2000; Trogkanis & Elkan, 2010).

Além disso, é aconselhável evitar a separação de dissílabos de quatro letras (por exemplo, *para*, *como*, *cede*). Essas considerações também levam a um texto mais esteticamente agradável e inteligível e ao controle do TEX sobre fragmentos isolados, por meio das variáveis `\lefthyphenmin` e `\righthyphenmin`.

Em algumas situações, o hífen deve ser repetido no início da linha seguinte. São elas:

1. casos em que palavras compostas ligadas por hífen são separadas entre linhas (por exemplo, *couve-/-flor*, *ex-/-presidente*); e
2. casos em que a separação de um pronome possa resultar em alteração de sentido (por exemplo, *prazer de ver-/-me*).

Sistematizar as regras que orientam os limites silábicos e a hifenização do português é um passo fundamental para entender e melhorar as regras de hifenização do TEX. Para garantir a precisão das regras de hifenização e comparar seus resultados de forma eficaz, um dicionário ortográfico de hifenização funciona como uma referência indispensável. Esse dicionário fornece uma lista abrangente de palavras hifenizadas corretamente, permitindo o cálculo da exatidão de cada conjunto de regras. Ao consultar o dicionário ortográfico de hifenização, é possível verificar se

[7] É importante ressaltar que as formas *connosco* e *commummente* não são usadas em todas as variedades do português. Por exemplo, os europeus adotam essas formas, enquanto os brasileiros não.

os padrões de hifenização gerados por um conjunto de regras estão de acordo com os padrões estabelecidos. Esse processo envolve analisar o quão bem as regras de hifenização aderem às convenções aceitas do idioma, garantindo que elas segmentem palavras de forma eficaz, sem comprometer a legibilidade ou a consistência. Além disso, comparando os resultados obtidos de diferentes conjuntos de regras de hifenização com as entradas do dicionário ortográfico, podemos avaliar a eficácia e a confiabilidade de cada abordagem.

Na Seção 5, definiremos as especificações desse dicionário de referência, delineando seu papel em atualizações e geração de regras subsequentes.

5. Criando o dicionário de hifenização para referência

Nosso conjunto de palavras em português foi criado a partir do corpus CETENFolha (Linguateca, 2014), ampliado com a lista de palavras do Palavras NET (sd.) e com o dump da Wikipédia em português (Wikipedia, 2023). A partir dos dados da Wikipédia, reduzimos a seleção a um subconjunto de 50.721 palavras, que representam 95% das ocorrências no dump (o conjunto inicial possuía 419.578 palavras). Esse limite foi estabelecido para filtrar erros de digitação e palavras pouco frequentes. Por fim, refinamos a lista, mantendo apenas as palavras para as quais encontramos dados de hifenização em pelo menos um dos seguintes recursos lexicais online: Michaelis (sd.), Priberam (sd.), Wikcionário (sd.), Aulete (sd.), Portal da Língua Portuguesa (sd.) e Dicio (sd.). Esse processo de curadoria resultou em um dicionário final contendo 86.324 palavras. Também realizamos correções manuais em 10 entradas que julgamos necessárias.

O fato de a língua portuguesa ter a fonologia como fator-chave para a hifenização exige uma compreensão de como a acentuação influencia esse processo. Toda palavra no idioma com duas ou mais sílabas possui uma sílaba mais proeminente, considerada sílaba tônica[8]. A tonicidade pode recair sobre qualquer uma das três últimas sílabas da palavra, contadas a partir da margem direita. As palavras são classificadas como oxítonas quando a última sílaba é tônica, paroxítonas quando a penúltima sílaba é tônica, ou proparoxítonas quando a antepenúltima sílaba é tônica. O padrão de tonicidade considerado não marcado (o mais comum) não recebe marcação diacrítica; os padrões marcados, incluindo todas as proparoxítonas, recebem uma marcação diacrítica. Essa marcação serve para indicar ao leitor que a sílaba tônica segue um padrão menos comum (marcado). O uso de marcas diacríticas para indicar a tonicidade é uma pista importante para determinar como uma palavra pode ser hifenizada. Por exemplo, as palavras *saúde* e *saída*, que possuem a tonicidade marcada nas vogais *u* e *i*, respectivamente, indicam claramente que essas vogais formam sílabas próprias: *sa-ú-de* e *sa-í-da*. No entanto, existem casos ambíguos, tanto para falantes, quanto na literatura linguística, como *his-tó-ria* (ou *his-tó-ri-a*) e *car-tó-rio* (ou *car-tó-ri-o*). A questão é se grupos vocálicos finais como *-ia* e *-io* são ditongos ou hiatos; decisão essa que mudaria tanto a hifenização da palavra como a posição em que a sílaba tônica se encontra. Adotamos, nesta pesquisa, a posição de autores como Cunha & Cintra (2016); Cegalla (2020); Pestana (2022), os quais assumem que palavras que terminam em ditongos crescentes recebem acento gráfico e são, portanto, paroxítonas, embora se reconheça que podem ser também proparoxítonas aparentes Cunha & Cintra (2016); Cegalla (2020).

No Wikcionário português, encontramos 1.848 palavras marcadas como proparoxítonas aparentes. As proparoxítonas aparentes são palavras que parecem ter a sílaba tônica na antepenúltima sílaba devido a sequências vocálicas pós-tônicas tratadas como ditongos crescentes. No entanto, para serem verdadeiras proparoxítonas, essas sequências devem formar um hiato; se forem ditongos, as palavras são classificadas como paroxítonas. Apesar da aparência, elas seguem as regras de acentuação das paroxítonas. As proparoxítonas aparentes podem gerar confusão na silabação e na hifenização porque seu padrão de tonicidade sugere diferentes pontos de quebra. Por exemplo, a palavra *início* poderia ser hifenizada de duas maneiras: *i-ní-ci-o* (proparoxítona) ou *i-ní-cio* (paroxítona). Essa duplicidade surge do tratamento das sequências vocálicas finais como sílabas separadas ou como partes de ditongos. Do ponto de vista linguístico, a silabificação acertada deve considerar a pronúncia pretendida e a estrutura morfológica, alinhando-se com as regras de acentuação padrão do idioma (Senado Federal, 2013).

A fim de analisarmos o desempenho das regras de hifenização e de propormos melhorias, extraímos quatro sub dicionários do dicionário principal: um primeiro dicionário em que todos dicionários online compartilham uma mesma

[8] É preciso destacar, entretanto, que há na língua portuguesa monossílabos tônicos, como *pá*, *pé* e *pó*. Não a incluímos, tendo em vista que elas não estão dentro do escopo da presente pesquisa.

hifenização (dicionário 6, com 15.842 palavras); um segundo dicionários em que cinco dicionários online propõem uma mesma hifenização (dicionário 5, com 15.642 palavras); um terceiro dicionário em que quatro dicionários online utilizam uma mesma hifenização (dicionário 4, com 10.299 palavras); e, por fim, um quarto dicionário em que três dicionário online propõem uma mesma hifenização (dicionário 3, com 7.745 palavras).

Em caso de desacordo entre diferentes hifenizações propostas pelos dicionários online, devemos ter cautela e examinar cada caso individualmente. Poderíamos ficar tentados a escolher a abordagem de maioria simples, selecionando apenas as hifenizações que reúnam mais defensores. No entanto, não está claro como as hifenizações nesses dicionários online foram curadas; muitos podem usar abordagens algorítmicas, levando a possíveis falhas compartilhadas entre eles. Por exemplo, cinco dicionários usam a hifenização *quart-zo*, enquanto o dicionário Aulete (originário de Portugal) favorece a hifenização *quar-tzo*, o que pode ser explicado pela ausência de vogal epentética no português europeu[9], conforme discutido em Collischonn (1999), que menciona Mateus (sd.). Similarmente, *ap-nei-a* ou *a-pnei-a* e *disp-nei-a* ou *dis-pnei-a* devem ser aceitas considerando que pode haver uma vogal epentética (português brasileiro) ou não (português europeu). Justificativa análoga pode ser atribuída a *hiperalgesia*, já mencionada na Seção 4.1. Dentre outros exemplos, podemos destacar também *neerlandês*, que aparece hifenizado como *ne-er-lan-dês* em cinco dicionários e como *neer-lan-dês* em um. Se considerarmos a pronúncia, preferiríamos a primeira hifenização, mas a segunda se conforma melhor às sílabas escritas regulares em português. Algumas sequências podem aceitar mais de uma hifenização, pois representam palavras diferentes. Por exemplo, *sub-li-nha* e *su-bli-nha*, onde a primeira é o substantivo *sublinha* e a segunda é uma forma flexionada do verbo *sublinhar*.

As proparoxítonas aparentes são casos de ambiguidade em que alguns dicionários optam por um hiato final (proparoxítonas), enquanto outros optam pelo ditongo final (paroxítonas) e alguns preferem marcar a possibilidade de existir ou não a separação entre as duas vogais finais. Para lidar com esses casos[10], utilizamos o seguinte procedimento: a) se a palavra tiver acento gráfico, selecionamos a opção paroxítona (por exemplo, *po-lí-cia* em vez de *po-lí-ci-a* e *pe-rí-neo* em vez de *pe-rí-ne-o*); b) se não tiver acento gráfico, escolhemos a opção proparoxítona (por exemplo, *au-to-ri-a* em vez de *au-to-ria* e *sim-bo-lo-gi-a* em vez de *sim-bo-lo-gia*)[11].

6. Regras de hifenização para o português

de Rezende (1987) criou os padrões iniciais para hifenização do português no TEX. O conjunto de padrões foi atualizado pela última vez em 2015, incorporando contribuições de de Rezende & Almeida (2015), resultando em um total de 307 regras.[12] Essas regras hifenizam efetivamente a maioria das palavras portuguesas.

Dentre as regras padrão do TEX, 252 (82%) seguem o padrão `1cV`, que representa as sílabas CV recorrentes em português. No que diz respeito às consoantes escritas, existem tipicamente 18 consideradas (`b`, `c`, `ç`, `d`, `f`, `g`, `j`, `k`, `l`, `m`, `n`, `p`, `r`, `s`, `t`, `v`, `x`, e `z`) que podem ser combinadas com 14 vogais escritas (`a`, `e`, `i`, `o`, `u`, `á`, `â`, `ã`, `é`, `í`, `ó`, `ú`, `ê`, e `õ`). Essas combinações resultam em padrões CV que indicam pontos de hifenização favoráveis antes das consoantes[13]. Vale ressaltar que pode haver outras regras ou exceções nos padrões de hifenização não cobertas por essas regras padrão. Para acomodar exceções, as seguintes regras foram propostas por de Rezende (1987):

- 20 regras foram criadas para casos envolvendo consoantes `b`, `c`, `d`, `f`, `g`, `k`, `p`, `t`, `v` ou `w`[14] seguidas de `l` ou `r`;

[9] O dicionário online Aulete tem origem na versão impressa conhecida como *Dicionário Caldas Aulete*. É um dicionário de Portugal, refletindo a variedade do português europeu.

[10] Em termos de translineação, devemos aplicar regras que impeçam uma vogal isolada na linha seguinte, entretanto, como mencionamos anteriormente, o TEX já faz este controle. O que queremos aqui é escolher uma forma de separação silábica que iremos adotar para ser base para as regras de hifenização.

[11] Apenas precisamos ter certeza de não selecionar aqueles casos em que há uma sequência da forma `[gq]u-[aeio]` (usando notação regex), como em *a-pa-zi-gu-ar* e *en-xa-gu-ar*.

[12] O conjunto de regras para hifenização do português é pequeno em comparação com regras de outros idiomas. Por exemplo, o inglês atualmente possui 4.938 regras, o russo possui 7.023 regras e o alemão possui 34.011 regras.

[13] Em português, geralmente a letra *h* aparece na posição inicial da palavra ou antes das consoantes *c*, *l* ou *n*. Portanto, não é interessante incentivar um ponto de hifenização antes de um *h*V. Exemplos de palavras que possuem um *h* no meio e não são precedidas por *c*, *l* ou *n* são: *Bahia*, *Corinthians*, *show*, *shopping*, *Matheus* e *sushi*. Esses são nomes próprios ou palavras estrangeiras cuja grafia não foi adaptada ao português.

[14] Não se sabe ao certo por que adicionaram as regras `1w2l` e `1w2r`, já que não há palavras no português com essa sequência. Encontramos apenas palavras estrangeiras como *showroom*, *wrestling*, *bowling* ou outras palavras estrangeiras bem raras.

- 3 regras para `c`, `l` ou `n` seguidas de `h`;
- 23 padrões distintos foram introduzidos para indicar pontos de hifenização entre vogais ou entre `c`'s, `r`'s e `s`'s;
- 8 padrões seguem o padrão `1[gq]u4V`, sinalizando pontos de hifenização favoráveis antes de `g` ou `q`, seguidos por uma sequência de `u` e uma vogal, com um ponto de inibição entre o `u` e a vogal subsequente;
- 1 padrão representado como `1-`, indicando que um hífen realmente funciona como um ponto de hifenização favorável.

No entanto, diante de certos casos, uma análise das regras padrão foi realizada para identificar áreas para melhoria (Araujo & Benevides, 2024). Ao lidar com problemas específicos e considerar padrões atípicos, nosso objetivo foi aumentar a precisão da hifenização. A metodologia utilizada para esta análise será descrita a seguir.

6.1. Atualizações para as regras de hifenização do português

O primeiro passo para propor atualizações ao conjunto de regras de hifenização é avaliar o desempenho delas em nosso dicionário de hifenização de referência. A Tabela 1 resume os resultados, mostrando o número de palavras corretamente hifenizadas pelas regras padrão, o número de palavras hifenizadas incorretamente e o número de palavras em que um ponto de hifenização foi omitido.

dicionário	correto	incorreto	faltante	palavras
dic. 6	15537	30	277	15842
dic. 5	13981	1146	601	15642
dic. 4	9001	883	518	10299
total	38519	2059	1396	41783

A soma das hifenizações corretas com as incorretas e as faltantes pode ser maior que o número total de entradas, pois alguns casos podem conter hifenizações incorretas e faltantes ao mesmo tempo.

Tabela 1: Resultados da aplicação das regras padrão de hifenização do TEX aos dicionários 6, 5 e 4.

Para criar regras complementares que resolvam os casos de hifenizações incorretas e faltantes, adotamos a seguinte abordagem: começamos resolvendo os casos de hifenizações incorretas e possíveis exceções necessárias para as regras propostas; em seguida, buscamos resolver os casos de hifenizações faltantes, também analisando a necessidade de criarmos regras de exceção. Este processo foi realizado interativamente, começando pelo dicionário 6, passando ao dicionário 5 e terminando no dicionário 4. O dicionário 3 foi reservado para avaliar a generalização das regras propostas.

Abaixo, sistematizamos o conjunto de regras adicionais propostas e fornecemos alguns exemplos de aplicação para cada uma delas. Este conjunto de regras deve ser utilizado de forma adicional às regras propostas por de Rezende & Almeida (2015).

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*
exceção: `.su3b4li` — *sublinhar*, *sublimar*

12 `.ne4o` — *neoliberal*, *neonazista*

13 `1p2seu1d` — *pseudônimo*

14 `1qu` — *enquanto*, *inquieto*, *farquhar*, *quiloqubit*

15 `a1ir.`, `u1ir.` — *sair*, *extrair*, *diminuir*, *incluir*

16 `a1ind`, `a1i1nh` — *ainda*, *rainha*

17 `e1imp` — *reimpresso*, *teleimpressor*

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

19 `u1iz.`, `a1iz.` — *juiz*, *raiz*

20 `pro1i1b` — *proibição*

21 `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*
exceção: `tu2id`, `tu2it`, `co2ima`, `o2i1na` — *gratuidade*, *intuito*, *coima*, *boina*

22 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*, *preênsil*, *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*
exceção: 1gu2á, 1gu2ã, 1gu2é, 1gu2ê, 1gu2í, 1qu2á, 1qu2ã, 1qu2â, 1qu2é, 1qu2ê, 1qu2í, — *jaraguá*, *saguão*, *alguém*, *português*, *linguística*, *aquático*, *camaquã*, *equânime*, *inquérito*, *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*

O resultado da hifenização do novo conjunto de regras é apresentado na Tabela 2. Observamos uma redução significativa do número de hifenizações incorretas, entretanto, há aumento no número de casos em que pontos de hifenização não foram identificados. Isto ocorre devido às regras 4a., 4e. e 4o. (sugeridas em 24).

Ao analisarmos o dicionário 5, o número de hifenizações incorretas aumentou substancialmente. A maioria desses erros decorreu de proparoxítonas aparentes. Esses casos foram facilmente corrigidos evitando o hífen final que deixa as vogais *a*, *e* ou *o* sozinhas na última sílaba. Para tanto, foram introduzidas essas regras em 24. Embora um conjunto de regras mais abrangente pudesse ser desenvolvido para levar em conta todos os cenários, optamos por regras mais concisas, apesar de suas limitações. A adição delas reduz o número de hifenizações incorretas; no entanto, aumenta o número de hifenizações ausentes. A grande maioria dessas hifenizações ausentes adicionais ocorre antes da vogal final de uma palavra. Considerando o menor número de casos e a posição final (antes de uma vogal final), a gravidade percebida (uma hifenização ausente é considerada menos problemática do que uma incorreta, sobretudo em final de palavra) é insignificante; portanto, optamos por mantê-las.

dicionário	correto	incorreto	faltante	palavras
dic. 6	15304	0	538	15842
dic. 5	14814	20	818	15642
dic. 4	9690	13	603	10299
total	39808	33	1959	41783

A soma das hifenizações corretas com as incorretas e as faltantes pode ser maior que o número total de entradas, pois alguns casos podem conter hifenizações incorretas e faltantes ao mesmo tempo.

Tabela 2: Resultados da aplicação dos novo conjunto de regras na hifenização das palavras dos dicionários 6, 5 e 4.

7. Elaboração de novas regras usando o *patgen*

Em Araujo & Benevides (2024), também abordamos a criação de novas regras utilizando o *patgen*. Este processo envolve a definição de parâmetros específicos, a escolha de um dicionário de referência com hifenizações corretas e a utilização de um arquivo de tradução personalizado para o *patgen*. Na Seção 3.1, detalhamos como o *patgen* funciona, quando mencionamos a necessidade de utilizarmos um arquivo de tradução para a língua em questão. As primeiras 10 linhas do arquivo de tradução utilizado para o português é apresentado na Listagem 3.

Listagem 3: 10 primeiras linhas do arquivo de tradução para o português.

```
 1 1
%% This file portuguese.tra defines the
    letters used for generating
%% Portuguese hyphenation patterns with
    patgen.
 a A
 á Á
 à À
 â Â
 ã Ã
 b B
 c C
```

Também mencionamos anteriormente que o *patgen* necessita de um dicionário de hifenização de referência, que será sua base de dados para extrair padrões. O dicionário de referência utilizado foi a junção dos dicionários 6, 5 e 4, sendo que o 3 foi reservado para testar o conjunto final de regras elaboradas.

Ainda, como destacado anteriormente na Seção 3.1, precisamos definir certos parâmetros numéricos na execução do *patgen*. Existem infinitas maneiras de especificar tais parâmetros. Com base em trabalhos anteriores, optamos por testar duas abordagens: (1) utilização de parâmetros fixos em todas as etapas da execução do *patgen* (partindo de um conjunto vazio ou das regras

padrão) (Haralambous, 2021); e (2) utilização de uma abordagem progressiva, começando com um conjunto vazio e criando regras de ordem superior a cada etapa, sempre com base nas regras da etapa anterior (Sojka & Sojka, 2019).

Embora os resultados do conjunto de regras geradas pelo *patgen* pareçam excepcionais no conjunto de palavras usadas para criá-las, devemos ser céticos em relação à sua generalização. Para investigar esse assunto, utilizamos essas regras no dicionário 3. Como demonstrado em Araujo & Benevides (2024), o *patgen* cria um grande conjunto de regras, muitas delas longas e específicas, resultando em baixa capacidade de generalização e desempenho inferior às regras manuais propostas no dicionário 3. Resultados semelhantes foram observados quando o *patgen* começa a partir do conjunto padrão de regras ou do nosso conjunto de regras corrigidas.

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

Tabela 3: Hifenização de palavras evidenciando o conflito entre o critério morfológico e o critério fonológico na hifenização utilizada nos dicionários.

8. Limitações

De forma geral, embora as palavras incorretas compartilhem algumas características comuns que poderiam permitir, até certo ponto, a redução de erros de hifenização, existe uma limitação inerente à maneira como as regras do TEX são concebidas. Abaixo, apresentamos os padrões encontrados nesses dados que poderiam ser contemplados por uma estrutura de regras diferente.

Determinação Morfológica Prefixos como *re-*, *sub-*, *ciber-*, *hiper-* e *auto-*, entre outros, exigem a separação do prefixo de seu radical, o que pode levar a questões fonológicas. Por exemplo, palavras como *reiniciar*, *sublinhar*, *ciberespaço*, *hiperalgesia* e *autoimagem* contêm prefixos e poderiam ser hifenizadas como *re-i-ni-ci-ar*, *sub-li-nhar*, *ci-ber-es-pa-ço*, *hi-per-al-ge-si-a* e *au-to-i-ma-gem*, respectivamente, para respeitar sua formação morfológica. No entanto, considerando que o português hifeniza suas palavras com base em correlatos fonológicos silábicos, palavras que requerem informação morfológica, como essas, podem não ter sua hifenização realizada corretamente. Estes exemplos são apresentados na Tabela 3, comparando-se a forma como diferentes dicionários as hifenizam.

Estrangeirismos Um grupo de palavras do corpus é formado por terminologias ou vocábulos incorporados ao português sem completa adaptação fonológica, como *darwinismo*, *quilowatt* e *esfiha*. Essa falta de adaptação resulta em padrões fonológicos altamente específicos para cada palavra, tornando impraticável integrá-las às regras do TEX ao lado de outras. Podemos, entretanto, adicioná-las à lista de exceções de hifenização. Para palavras como *farquhar* e *qubit*, que são estrangeiras, mas, apesar de raras, possuem frequência significativa, adicionamos a regra 14, já que tal regra de hifenização é coerente com as normas da língua portuguesa e não há necessidade de criar qualquer regra de exceção.

Grupos consonantais iniciais O português possui poucos casos de grupos consonantais no início de uma palavra. Geralmente, são resquícios etimológicos e, atualmente, improdutivos na língua, já que não há neologismos com esse padrão. Encontros como *ps-* e *pn-* são mais frequentes, pois estão presentes em palavras como *psicologia* e *pneu*, de média frequência na língua. Esses casos podem ser previstos por regras específicas que cobririam 49 e 13 palavras no corpus, respectivamente. No entanto, existem grupos consonantais encontrados em palavras bem específicas e de baixa frequência. Embora possível, não vale a pena adicionar regras muito específicas para grupos encontrados em palavras como *dzeta*, *gnu*, *cnidário*, *ftálico* e *gnaisse* – que representam apenas cinco palavras.

Abreviações, Acrônimos e Siglas Seja por eficiência, conveniência, clareza ou jargão especializado, é comum usar versões abreviadas de palavras ou frases. A abreviação

é um método empregado para alcançar encurtamento. Em nosso corpus de português, encontramos exemplos como *etc.*, *Dr.*, *Exmo.*, *cap.*, *Univ.*, *ed.*, *s.n.*. Outra forma abreviada é a sigla, que consiste em usar as letras iniciais das palavras para criar uma versão reduzida. No entanto, as siglas nem sempre seguem as regras de hifenização descritas neste trabalho, pois não necessariamente seguem os padrões ortográficos ou fonotáticos da língua. Algumas siglas encontradas no corpus incluem *SESC*, *INSS*, *PCdoB*, *PM* e *UFRJ*. Os acrônimos são um tipo específico de abreviação em que as primeiras letras (ou grupos de letras) de cada palavra são combinadas para formar uma nova palavra pronunciável. No corpus, encontramos exemplos como *Anatel*, *Ovni*, *Sida* e *Mercosul*. Essas várias formas abreviadas desempenham um papel importante na linguagem escrita, fornecendo maneiras concisas de representar palavras ou frases mais longas. É importante observar que abreviações, acrônimos e siglas geralmente são tratados como unidades únicas e não são hifenizados.

9. Conclusão

Neste estudo, exploramos as regras padrão de hifenização do TEX e analisamos as regras adicionais para melhorar o desempenho da hifenização em português. Ao abordar as limitações das regras antigas e das abordagens existentes no TEX, e ao incorporar considerações morfológicas e fonológicas, conseguimos reduzir significativamente o número de erros de hifenização. Contudo, o conjunto de regras adicional ainda não é suficiente para alcançar uma precisão perfeita, o que nem mesmo é desejado, pois nosso conjunto de dados pode conter ruído e há muitos casos de hifenização duvidosa. Optamos por um conjunto conciso de regras que possa generalizar melhor e, portanto, se alinhar adequadamente com as regras subjacentes de hifenização do idioma.

Também testamos regras criadas pelo *patgen*, que geraram um extenso conjunto de regras incapazes de generalizar de forma eficaz. Em contraste, nossas regras manuais apresentaram desempenho superior em um conjunto de validação.

Diante disso, embora nosso conjunto de regras aprimorado demonstre melhorias significativas na precisão da hifenização, ainda há espaço para refinamentos adicionais. Aspectos como o suporte a caracteres especiais e a segmentação silábica universal, conforme considerado por Sojka et al. (2023), merecem consideração. A incorporação de expressões regulares, a criação de classes de caracteres e o suporte à posição de acento tônico das palavras podem aumentar ainda mais a eficiência e a generalidade do sistema de hifenização. Tais aprimoramentos não apenas aumentariam a precisão, mas também tornariam o sistema mais robusto e adaptável às diversas nuances do português.

Referências

de Araújo, Antonio Martins & Toru Maruyama. 2015. A hifenização em português. *Idioma* 28. 90–107.

Araujo, Leonardo & Aline Benevides. 2024. Enhancing TEX hyphenation rules for Portuguese. *TUGboat* 45(3). 309–316. 10.47397/tb/45-3/tb141araujo-pthyph

Aulete. sd. Dicionário Aulete digital. Accessed: 2023-06-26.

Bergström, Magnus & Neves Reis. 2011. *Prontuário ortográfico e guia da língua portuguesa*. Casa das Letras

Cagliari, Luiz Carlos. 2015. Aspectos teóricos da ortografia. Em Maurício Silva (ed.), *Ortografia da língua portuguesa: história, discurso, representações*, 17–52. Contexto

Cegalla, Domingos Paschoal. 2020. *Novíssima gramática da línngua portuguesa*. São Paulo: Companhia Editora Nacional

Collischonn, Gisela. 1999. A epêntese e a fonologia lexical do português brasileiro. Em *14th Encontro Nacional da Associação Portuguesa de Linguística*, 369–382.

Coulmas, Florian. 2003. *Writing systems: An introduction to their linguistic analysis*. Cambridge University Press

Cunha, Celso & Lindley Cintra. 2016. *Nova gramática do português contemporâneo*. Lexikon Editora Digital 7th edn.

Dicio. sd. Dicionário online de Português. Accessed: 2023-06-26.

Gândavo, Pêro de Magalhães. 1981. *Regras que ensinam a maneira de escrever e a ortografia da língua Portuguesa*. Lisboa: Biblioteca Nacional.

Haralambous, Yannis. 2021. A revisited small tutorial on Patgen, 28 years after. Relatório técnico. CTAN.

Honorof, Douglas & Laurie Feldman. 2006. The chinese character in psycholinguistic research: Form, structure, and the reader. Em Ping Li, Li Hai Tan, Elizabeth Bates & Ovid J. L. Tzeng (eds.), *The Handbook of East Asian Psycholinguistics*, vol. 1, chap. 17, 195–208. Cambridge University Press. 10.2277/0521833337

Hunspell's Team. 2023. Hunspell. [Visitado a 2023-04-13].

Hunspell's Team. sd. Hunspell Hyphen.

Knuth, Donald Ervin. 1977. Preliminary preliminary description of TEX. Online.

Levien, Raph. 1998. *Brief explanation of the hyphenation algorithm herein*. Hunspell.

Liang, Frank & Peter Breitenlohner. 1991. PATtern GENeration program for the TEX82 hyphenator. Relatório Técnico. 2 CTAN

Liang, Franklin Mark. 1983. *Word hy-phen-a-tion by com-put-er*: Stanford University. Tese de Doutoramento

Libossek, Marion & Florian Schiel. 2000. Syllable-based text-to-phoneme conversion for German. Em *6th International Conference on Spoken Language Processing (ICSLP)*, 283–286.

Lin, Li-chin. 2011. Fundamental generalizations of English syllabification. *Concentric: Studies in Linguistics* 37(2). 179–208.

Linguateca. 2014. CETENFolha corpus. Accessed: 2023-06-26.

Mateus, Maria Helena Mira. sd. Questões fonológicas do português. Manuscrito

Michaelis. sd. Michaelis Dicionário brasileiro da língua Portuguesa. Accessed: 2023-06-26.

Németh, László. 2006. Automatic non-standard hyphenation in OpenOffice.org. *TUGboat* 27(1). 32–37

Palavras NET. sd. Palavras NET word list. Accessed: 2023-06-26.

Palmer, David D. 2010. Tokenisation and sentence segmentation. Em Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of Natural Language Processing*, chap. 2, 9–30. CRC Press

TEX pattern authors. sd. TEX hyphenation patterns.

Pestana, Fernando. 2022. *A gramática para concursos públicos*. Método

Portal da Língua Portuguesa. sd. Portal da língua Portuguesa online dictionary. Accessed: 2023-06-26.

Priberam. sd. Dicionário priberam da língua Portuguesa. Accessed: 2023-06-26.

de Rezende, Pedro Jussieu. 1987. Portuguese hyphenation table for TEX. *TUGboat* 8(2). 102–102

de Rezende, Pedro Jussieu & José Joao Dias Almeida. 2015. Hyphenation patterns for Portuguese.

Scannell, Kevin. 2003. Hyphenation patterns for minority languages. *TUGboat* 24(2). 236–239

Senado Federal. 2013. *Acordo ortográfico da língua portuguesa: atos internacionais e normas correlatas*. Senado Federal

Sojka, Ondřej, Petr Sojka & Jakub Máca. 2023. A roadmap for universal syllabic segmentation. *TUGboat* 44(2). 289–296. 10.47397/tb/44-2/tb137sojka-syllabic

Sojka, Petr. 1995a. Notes on compound word hyphenation in TEX. *TUGboat* 16(3). 290–297

Sojka, Petr. 1995b. Notes on compound word hyphenation in TEX. Relatório técnico. Masaryk University in Brno, Faculty of Informatics

Sojka, Petr. 2002. Hyphenation – a tutorial for TEX users.

Sojka, Petr & David Antoš. 2003. Context sensitive pattern based segmentation: A Thai challenge. Em *EACL Workshop on Computational Linguistics for South Asian Languages–Expanding Synergies with Europe, Budapest*, 65–72

Sojka, Petr & Ondřej Sojka. 2019. The unreasonable effectiveness of pattern generation. *TUGboat* 40(2). 187–193.

Sojka, Petr et al. 2005. *Competing patterns in language engineering and computer typesetting*: Masaryk University, Brno. Tese de Doutoramento

Trogkanis, Nikolaos & Charles Elkan. 2010. Conditional random fields for word hyphenation. Em *48th Meeting of the Association for Computational Linguistics (ACL)*, 366–374.

Wikcionário. sd. Wikcionário online dictionary. Accessed: 2023-06-26.

Wikipedia. 2023. Portuguese wikipedia dump. Accessed: 2023-06-26.

Yavas, Mehmet. 2020. *Applied English phonology*. John Wiley & Sons