DOI: 10.14393/DL30-v11n3a2017-13

Conversor de transcrição fonética automática para as formas linguísticas da variedade linguística potiguar

A converter of automatic phonetic transcription for the linguistic forms of the potiguar linguistic variety

Cid Ivan da Costa Carvalho*

RESUMO: Os conversores de grafema para fonema executam o processo de transcrição das formas gráficas de uma língua para as formas fonéticas. Esse trabalho apresenta o Potigrafone - conversor grafofônico para a variedade linguística potiguar. O sistema computacional Potigrafone se baseia em dois importantes atlas linguísticos, os quais apresentam as transcrições fonéticas da pronúncia potiguar: o *Atlas Linguístico do Centro-Oeste Potiguar* e o *Atlas Geolinguístico do Litoral Potiguar* (AliPTG). Para o desenvolvimento desse sistema, fizemos a implementação das regras da fonologia gerativa, formando a relação de um conjunto de representações subjacente com as representações de superfície. Esse sistema foi aplicado a essa variedade linguística, mas, para alguns fenômenos, pode ser aplicada a outras variedades linguísticas do país.

PALAVRAS-CHAVE: Fonética. Transcrição fonética automática. Variedade linguística.

ABSTRACT: A converter of grapheme to phoneme execute the transcription process of the grapheme forms of a language for the phonetic forms. This paper introduces the Potigrafone – a converter of grapheme to phoneme for the potiguar linguistic variety. Potigrafone computational system fundaments if in two linguistics atlas important, where they produce the phonetic transcriptions of the potiguar pronunciation: the *Atlas Linguístico do Centro-Oeste Potiguar* and the *Atlas Geolinguístico do Litoral Potiguar* (AliPTG). For the development this system, we make the implementation of the rules of the generative phonology, modeling the relation of a set underlying forms with the surface forms. This sistem it was applied for this linguistc variety, but can to be applied in other linguist variety of the country.

KEYWORDS: Phonetics. Automatic phonetic transcription. Potiguar linguistic variety.

1. Introdução

O uso de ferramentas computacionais como instrumento de auxílio às pesquisas linguísticas se torna um imperativo quando se trata de manipular extenso banco de dados. Nesse trabalho, apresentamos um sistema computacional que torna possível a transcrição de grande quantidade de material escrito para a representação fonética da fala potiguar. É claro que a transcrição manual parece ser mais precisa do que a transcrição automática, no entanto, este

* Professor Doutor da Universidade Federal Rural do Semiárido – UFERSA - e coordenador do Grupo de Estudo em Linguística Computacional – GELC.

tipo de transcrição apresenta erros sistemáticos e pontuais, os quais podem ser corrigidos com maior facilidade pelo usuário, diferentemente, do primeiro. Além disso, a transcrição é feita com menor tempo quando é feita automaticamente.

Quando se trata de conversão automática de grafema para fonema, os sistemas são chamados de *Grapheme to Phoneme* (G_2P). O Grafone[1] é um exemplo de sistema G_2P para o português e foi desenvolvido pelo Laboratório de Processamento de Sinais, em Coimbra. E os sistemas que fazem a conversão automática de grafema para fone são conhecidos como *Letters to Sounds* (L_2S). O Petrus[2] é um exemplo desse tipo de sistema e ele faz a transcrição do grafema do português para a variedade linguística paulista. Esse sistema foi desenvolvido pelo Núcleo Interinstitucional de Linguística Computacional (NILC).

Neste trabalho, temos o objetivo de apresentar o Potigrafone, sistema que executa o processo de conversão das formas gráficas do português para as formas fonéticas[3] da variedade linguística potiguar, ou seja, um L_2S que foi desenvolvido para essa variedade linguística brasileira. Destacamos três fatores relevantes presentes no seu desenvolvimento: (1) a inexistência de sistema automático para a variedade potiguar; (2) a presença de dados empíricos que apresentam os fenômenos fonéticos presentes na fala potiguar e que deram maior consistência à transcrição fonética do sistema e (3) muitos fenômenos fonéticos mostrados nas transcrições desse sistema estão presentes em outras variedades linguísticas do país, logo, ele pode auxiliar na pesquisa de outras variedades linguísticas.

O primeiro fator aponta para o fato de que a forma fonética desse sistema representa o registro do português falado Rio Grande do Norte, mais especificamente, em duas regiões que compreendem a maior parte do estado: as regiões Oeste e Litorânea. O segundo mostra que o desenvolvimento desse sistema recebeu contribuição empírica dos atlas vinculados ao projeto Atlas Linguístico do Brasil (AliB) que foram: Atlas Linguístico do Centro-Oeste Potiguar e do Atlas Geolinguístico do Litoral Potiguar (ALiPTG), ou seja, a transcrição fonética automática reflete os dados empíricos presentes nesses atlas. Por último, muitos fenômenos fonéticos apresentados na transcrição automática desse sistema estão presentes em outras variedades

[1] O Grafone é conversor de grafemas para fonemas para o Português Europeu e está disponível em: http://www.co.it.pt/~labfala/g2p/.
[2] O Petrus é um sistema Web de suporte à transcrição fonética automática do Português Brasileiro e está disponível em: http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources.
[3] Destacamos que a expressão "forma gráfica" diz respeito à escrita ortográfica da palavra na língua portuguesa e a "forma fonética" corresponde à transcrição fonética da palavra. No texto, a primeira é escrita entre colchetes angulares <> e a segunda entre colchetes [ ]. Assim, o símbolo <g> equivale à letra "gê" do alfabeto português e o símbolo [g] corresponde ao som oclusivo velar sonoro.

linguísticas do Brasil. Citemos, por exemplo, o fenômeno da nasalidade que ocorre na maioria das variedades linguísticas faladas no nordeste. Esse fato mostra que o sistema pode ser utilizado para outras variedades, mas o usuário deve atentar para as formas fonéticas apresentadas pelo sistema e verificar se realmente corresponde à representação da sua variedade linguística.

Pensando assim, dividimos este trabalho em três seções: a primeira apresenta algumas diferenças entre a transcrição automática do Petrus e do Potigrafone; a segunda mostra dois atlas linguísticos utilizados como base empírica para a compreensão das variantes linguísticas presentes na fala potiguar para a aplicação no sistema computacional; a terceira descreve e aplica dois conceitos da fonologia gerativa que fundamentam a implementação do Potigrafone: as regras fonológicas e os componentes fonológicos - a representação subjacente e a representação de superfície.

2. Transcrição fonética automática

A transcrição fonética é a representação dos sons emitidos por falantes de uma língua. A representação do som é feita por meio do uso de um conjunto de símbolos de um alfabeto fonético. Há dois tipos de transcrição fonética: a transcrição ampla e a restrita.

> Na transcrição restrita (detalhada), todos os detalhes fonéticos, mesmo aqueles que podem ser previsíveis pelo contexto, incluindo propriedades secundárias, são consideradas, e na transcrição ampla (aproximada) são explícitas apenas os aspectos mais gerais dos segmentos. (SEARA; NUNES; VOLCÃO, 2015, p. 83).

As autoras exemplificam essa distinção com a palavra <quilo>, que pode ser transcrita de forma restrita [ˈkʲilʷʊ] ou de forma ampla [ˈkilʊ]. A primeira transcrição considera certas propriedades secundárias, detectadas em determinados segmentos previsíveis pelo ambiente a ser transcrito, como a palatalização da consoante velar diante de vogal alta anterior e o arredondamento da lateral diante de vogal alta posterior. A segunda não considera nenhum desses detalhes. Quanto mais fiel for à pronúncia, mais recursos fonéticos devem ser usados, exigindo-se maior número de sinais e diacríticos de um alfabeto.

Nesse contexto, apresentamos a diferença na transcrição fonética automática executada por dois sistemas diferentes: pelo Petrus e pelo Potigrafone. Tanto o primeiro quanto o segundo

transcrevem as formas gráficas do português para a forma fonética ampla, ou seja, sem marcar certas propriedades secundárias da fala.

2.1 Petrus

O sistema Petrus executa a transcrição fonética automática das sequências de grafemas do português para o dialeto paulista, utilizando os símbolos fonéticos do Alfabeto Fonético Internacional (AFI). Marquiafável e Zavaglia (2011, p. 156) explicam que "Por dialeto paulista, deve-se entender a fala de pessoas cultas oriundas do Estado de São Paulo" e acrescentam que não há qualquer sistema desse tipo para dialeto padrão de São Paulo. Além disso, é um sistema que dispõe de uma interface gráfica de fácil utilização, servindo de apoio à transcrição fonética automática de palavras isoladas ou de um conjunto de palavras.

Segundo Marquiafável, Bokan, Zavaglia (2014), o Petrus é composto por seis módulos. O primeiro é responsável pela transcrição fonética das palavras homógrafas heterofônicas, como nas palavras 'g[o]sto (substantivo) e 'g[O]sto (verbo). Esse recurso linguístico foi construído usando um dicionário com pares de palavras heterofônicas com quase 1.000 homográficos transcritos. O segundo módulo executa a identificação da presença de prefixo em uma palavra antes de transcrevê-la, uma vez que os prefixos em PB têm um comportamento específico no sentido de alternação vocálica, e o recurso linguístico utilizado foi um algoritmo baseado com 145 regras. O terceiro módulo identifica o diacrítico ['], a vogal acentuada na palavra que está sendo analisada, usando um algoritmo baseado em regras. O quarto módulo identifica e marca os limites de sílabas e usa um algoritmo com 25 regras baseados em *Hide Markov Model* (HMM). O quinto módulo executa a etiquetagem das palavras - *Part-de-Speech tagger* - usando UNITEX. No último módulo, toda essa informação linguística é utilizada para transcrição fonética das palavras escritas, por meio de um conjunto de 160 regras, sendo que 62 delas fazem a transcrição das consoantes e 98 fazem das vogais.

Com a utilização desses módulos, o Petrus executa a transcrição das palavras, a separação silábica e a classificação das palavras transcritas quanto à classe gramatical. No Quadro 1, a seguir, podemos ver um exemplo de transcrição automática da palavra <peso>.

Quadro 1 – Exemplo de transcrição fonética pelo sistema *Petrus*.

Grafema	Transcrição fonética
peso	[ˈpezʊ] Syllable boundaries: pe.so Part of speech: Substantivo
	[ˈpɛzʊ] Syllable boundaries: pe.so Part of speech: Verbo

Fonte: dados extraído de análise no programa Petrus.

O Petrus é um sistema muito adequado ao uso lexicográfico, uma vez que ele não executa apenas a transcriação fonética, mais também a separação silábica e a classificação gramatical das palavras. No entanto, o conversor ainda comete erros básicos de silabação, como na palavra <ca.rro>, onde as letras dobradas ficam na mesma sílaba e também na transcrição fonética feita por esse sistema destaca apenas dois fenômenos fonéticos: a alternância vocálica, como foi o caso da palavra <peso>, e da epêntese, como é o caso da palavra <pneu>, transcrita por [ˈpɪneʊ], acrescida de [ɪ] epentético e a palatalização das oclusivas /t/ e /d/ antes da vogal anterior /ɪ/, como nas palavras <tia> e <dia>. A forma fonética do Petrus deixa de lado outros fenômenos fonéticos muito relevantes para a transcriação fonética dessa variedade, como, por exemplo, o apagamento das semivogais [ɪ] e [ʊ], para a formação de monotongo que ocorre nas palavras <peixe> [ˈpeʃɪ] e <tesoura> [teˈzoɾa], respectivamente, como afirma Tós (2010).

2.2 Potigrafone

O Potigrafone[4] é um conversor de transcrição automática utilizando da tecnologia de estados finitos para executar a conversão das formas gráficas para as formas fonéticas da variedade potiguar. Essa tecnologia é construída por duas fitas: uma fita de entrada e outra de saída. Por meio da primeira, o sistema reconhece as cadeias de caracteres da língua portuguesa, ou seja, a forma gráfica (*input*), e, por meio da segunda, gera a forma fonética para a variedade linguística potiguar, isto é, a cadeia de saída (*output*).

Considerando a relação da entrada com a saída do sistema, destacamos três aspectos peculiaridades na transcrição fonética desse sistema: (1) o alfabeto fonético, (2) o alinhamento das formas gráficas com a forma fonética e (3) a transcrição de fenômenos fonéticos.

[4] Para maiores informações, acesse a tese de doutorado em Linguística - *Transdutor de estados finitos para conversão de grafema para a pronúncia da variedade linguística potiguar*, da Universidade Federal do Ceará.

As formas fonéticas geradas por esse conversor são constituídas pelos símbolos e pelos diacríticos do alfabeto fonético *Speech Assessment Methods Phonetic Alphabet* (SAMPA). Para quem está habituado aos símbolos do Alfabeto Fonéticos Internacional (AFI), os símbolos do SAMPA não se tornam estranhos devido a dois fatores básicos: (1) a transcrição fonética feita com a utilização do SAMPA considera que as letras minúsculas do AFI possuem a mesma representação e quando não têm qualquer relação, os diacríticos do SAMPA seguem os símbolos que os modificam, Segundo Wells (2014); (2) os símbolos desse alfabeto estão presentes nos teclados dos computadores, logo, é reconhecido pelos usuários e pelos módulos dos processadores modernos como os *smartphones* e *tabletes*.

Nesse alfabeto, há vários diacríticos disponíveis para a transcrição da fala humana. Todavia, o Potigrafone utiliza apenas três deles na transcrição da fala potiguar: o til (~) para marcar a nasalização, a aspa dupla (") para sinalizar o acento tônico primário e a aspa simples (') para apontar a palatalização, conforme descrição a seguir:

a) a nasalização vocálica é marcada por meio de um til (~) após o símbolo vocálico, como em <canto> [ka~tu] e <cama> [ka~ma], sendo que o primeiro marca a vogal nasal e o segundo destaca o fenômeno da nasalidade;
b) o acento tônico é marcado por meio do acréscimo da aspa dupla (") antes da sílaba tônica, como em <canto> ["ka~tu];
c) a palatalização é representada por meio do acréscimo de aspa simples ('), após consoante palatalizada como em <óleos> ["Ol'us].

Outro aspecto importante é o alinhamento das formas gráficas com a forma fonética. O Potigrafone mostra duas maneiras de fazer a correspondência dos grafemas para as variantes linguísticas, ou seja, o alinhamento de grafema para o fone. Beesley e Karttunen (2002) dizem que o alinhamento é o primeiro passo que se deve fazer no desenvolvimento de sistema como esse. Isso é feito através de uma associação de um símbolo gráfico da forma lexical para um símbolo fonético na forma fonética, de um para um. Alguns grafemas do português possuem correspondências unívocas com os sons da variedade linguística potiguar, situação na qual o sistema faz a conversão de modo direto. É o caso das letras <p, b, t, d, f, v>, que são convertidos pelos símbolos [p, b, t, d, f, v], respectivamente.

Porém, outros possuem correspondências diretas dependendo de fatores contextuais e posicionais. Podemos exemplificar esse tipo de correspondência por meio das letras <g> e <r>. O grafema <g> se realiza como uma fricativa alvéolo palatal sonora [Z] ou como uma oclusiva velar sonora [g], dependendo do contexto. A saída do grafema <r> é realizada como uma fricativa glotal surda [h] ou como uma vibrante simples [ɾ], dependendo da posição silábica em que a letra se encontra; quando está na posição de coda medial ou no ataque inicial de palavra, é uma consoante glotal e quando está posicionada como segunda consoante do ataque complexo ou quando está entre vogais é uma consoante vibrante.

No entanto, os dígrafos consonantais e vocálicos são um conjunto de grafemas para representar apenas um som da variedade linguística, ou seja, o alinhamento não foi feito por meio da correspondência direta do grafema para o fone. Nesse caso, aos dígrafos consonantais foi acrescentado um símbolo fonético nulo [0] que corresponde às letras diacríticas que vêm junto às consoantes. Para exemplificar essa forma de transcrição, mostramos a Figura 1, que ilustra a transcrição alinhada do dígrafo <ch>, todavia, mantemos o alinhamento também nos outros dígrafos como: <lh>, <nh>, <rr>, <ss>, <sc>, <sç>, <xc>, <gu> e <qu>.

Figura 1 – Alinhamento dos dígrafos consonantais Forma gráfica.

Forma gráfica	C	h	a	v	e	s
	|	|	|	|	|	|
Forma alinhada	“S	0	a	v	i	s

Fonte: elaborada pelo autor.

Além do alinhamento, consideramos que a transcrição fonética de alguns dígrafos depende do contexto de escrita e da construção silábica, mais precisamente, da fronteira silábica. Nas palavras <crescer> e <casca>, as letras <sc> formam um dígrafo e um encontro consonantal, respectivamente, sendo que a transcrição fonética da letra <c>, na primeira palavra, tem como valor fonético o símbolo nulo [0], uma vez que nesse caso essa letra corresponde a uma forma diacrítica e, na segunda palavra, a letra <c> tem valor fonético de oclusiva velar surda [k], ou seja, essa letra representa uma consoante em posição de ataque simples.

Já para os dígrafos vocálicos, não foi utilizado um símbolo fonético nulo, mas um símbolo diacrítico do alfabeto fonético SAMPA, o til [~], para representar a nasalização da

vogal. Na forma escrita do português, as letras diacríticas vocálicas <n> e <m>, em posição pós-vocálicas, representam a nasalização da vogal, como na palavra <canto> [ka~tu]. Além disso, esse símbolo também foi aplicado no fenômeno da nasalidade dos segmentos vocálicos, seguido de uma das duas consoantes nasais, [m] ou [n], como nas palavras <cama> [ka~ma] e <cana> [ka~na]. Assim, no primeiro caso, o sistema executa a transforma das letras pelo diacrítico do SAMPA e, no segundo, ele insere um segmento antes das consoantes nasais.

O terceiro aspecto diz respeito à transcrição de fenômenos fonéticos que estão presentes na fala potiguar. Aqui, mencionamos apenas os fenômenos fonéticos da harmonia vocálica, da lenição, da monotongação, da ditongação e da epêntese, como mostra Quadro 2. No entanto, remetemos o trabalho de Carvalho (2016) para o conhecimento de outros fenômenos e das regras utilizadas no desenvolvimento do sistema.

Quadro 2 – Fenômenos fonéticos com mais de um *output*.

	Fenômenos fonéticos				
	Harmonia vocálica	*Lenição*	*Monotongação*	*Ditongação*	*Epêntese*
Forma gráfica	cebola	mulher	caixa	paz	pneu
Formas fonéticas	[se"bol6] [si"bol6]	[mu"Le] [mu"le]	["kajS6] ["kaS6]	["pajs] ["pajs]	[pi"new] [pe"new]

Fonte: elaborado pelo autor.

Para esses fenômenos, a transcrição automática apresenta duas ou mais formas fonéticas para a mesma forma gráfica. Por exemplo, a palavra <caixa> apresenta duas formas fonéticas: uma que marca o fenômeno da monotongação, ["kaSa], e outra sem esse fenômeno, ["kajSa]. Essa opção está respaldada no fato de que, na fala potiguar, ocorrem essas duas variantes linguísticas da forma escrita. Essa forma de transcrição para essa palavra e para outros casos se fundamenta nas cartas fonéticas do Atlas Linguístico do Centro-Oeste Potiguar e do AliPTG. Nesses casos, o Potigrafone recebe como entrada (*input*) uma palavra escrita graficamente, conforme a ortografia oficial e retorna como saída (*output*) duas ou mais formas fonéticas da fala potiguar.

Todavia, nem todos os fenômenos presentes nas cartas fonéticas desses dois atlas foram considerados na implementação do sistema. O fenômeno do metaplasma por transposição, como ocorre na carta fonética da palavra <prateleira>, mostra quatro formas fonéticas diferentes. Nas cinquenta e seis entrevistas, ocorreram 37,5% para a pronúncia [paht"lera], 26,7% para a pronúncia [prati"lera], 23,3% para a pronúncia [pati"lera] e 5,4% para a pronúncia [prati"lejra]. Esse tipo de fenômeno e o apagamento da vogal nas palavras proparoxítonas, como na palavra <fósforo>, que é pronunciada ~~por~~ ["fOsfru], o sistema não apresenta a transcrição, mas exibe apenas duas saídas e uma delas, obrigatoriamente, representa a forma mais próxima da entrada, ou seja, a transcrição [prati"lejra], que foi a menos frequente nos corpora. No entanto, é a mais alinhada com a entrada da palavra.

Portanto, esse conversor reconhece as cadeias de caracteres da língua e gera a forma fonética para a variedade linguística potiguar, marcando na transcrição alguns fenômenos fonéticos que não estar presente no Petrus e faz a acentuação silábica. Os atlas linguísticos embasaram a marcação desses fenômenos para essa variedade linguística e contribuem para que a transcrição seja fiel à fala potiguar, como trataremos a seguir.

3. Atlas linguísticos da variedade linguística potiguar

A base empírica do sistema computacional Potigrafone parte de dois importantes atlas linguísticos, os quais apresentam as transcrições fonéticas da pronúncia potiguar: o Atlas Linguístico do Centro-Oeste Potiguar e o Atlas Geolinguístico do Litoral Potiguar (ALiPTG) que fazem parte do projeto do Atlas Linguístico do Brasil (ALiB[5]). No que se refere a esses atlas, a transcrição das cartas fonéticas foi feita por linguistas e os dados contribuem na identificação e na quantificação dos fenômenos linguísticos, para a descrição da maior parte dos fenômenos linguísticos do Estado.

Outra característica é o fato de que a recolha dos dados se processa mediante a utilização de ficha de identificação do informante com três questionários que seguem integralmente o modelo do Projeto ALiB. Os modelos usados na pesquisa são o questionário semântico-lexical (QSL), com 202 perguntas que cobrem 14 áreas semânticas; o questionário Fonético-fonológico (QFF), com 159 perguntas e o questionário morfossintático (QMS), com 49 perguntas. (PEREIRA, 2008).

[5] Para mais informações, acesse o sítio do ALIB: https://twiki.ufba.br/twiki/bin/view/Alib/WebHome.v

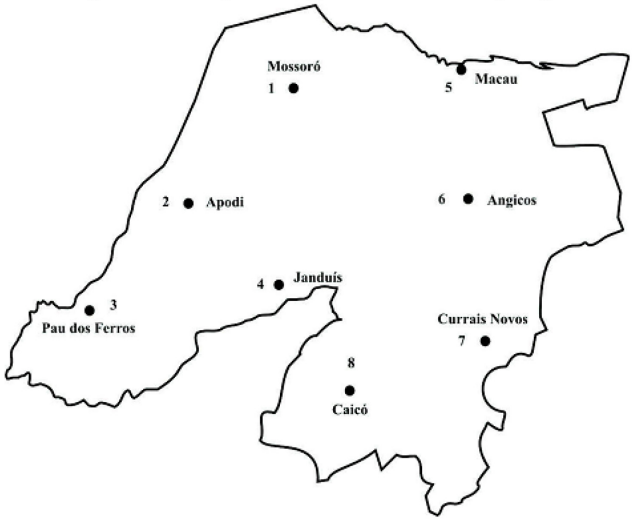
Todavia, foi dado enfoque aos fatos linguísticos apresentados nas cartas fonéticas dos atlas mencionados, pois essas cartas

> registram as variantes de um fonema comprovadas nos pontos investigados, ou de vários fonemas correspondentes a um único fonema mais antigo ou também determinadas séries de fonemas quase encontram na mesma situação do ponto de vista histórico. (SILVA, 2012, p. 110).

Ressaltamos que os atlas a seguir apresentam a diversidade de usos em sua distribuição espacial, com resultados acrescidos de algumas notas e ilustrações que complementam as informações sociolinguístico-culturais. Eles não contêm interpretação de dados, mas apenas análise linguística. Segundo Silva (2012)[6], o Atlas Linguístico do Centro-Oeste Potiguar foi elaborado a partir da identificação das variáveis extralinguísticas diastrática, diassexual e diageracional dos fenômenos fonéticos e lexicais e da descrição da realidade do português do Centro-Oeste Potiguar.

Considerando os dados socioeconômicos e culturais particulares da região pesquisada, esse autor selecionou oito pontos de inquéritos: quatro da mesorregião do oeste Potiguar (Mossoró, Apodi, Pau dos Ferros e Janduís) e quatro da mesorregião central Potiguar (Macau, Angicos, Currais Novos e Caicó), como mostra a Figura 2.

Figura 2 – Mapa da região centro-oeste potiguar.



Fonte: Silva (2012, p. 139).

[6] Neste subtópico, apresentamos as principais informações expostas por Silva (2012). Para mais informações sobre a pesquisa feita, a tese do autor deve ser consultada.

O autor acrescenta que o critério para a escolha dessas cidades foi feito com base na importância dos aspectos demográficos, históricos, geográficos, políticos, econômicos e culturais e na influência destes aspectos sobre os outros municípios da mesorregião. Adotou-se, também, o critério da equidistância aproximada. Assim, todas as localidades foram distribuídas de maneira que abrangessem todo o centro-oeste Potiguar, com uma distância entre elas, de, pelo menos, 70 km.

Para a realização da pesquisa de campo, feita por Silva (2012), foram selecionados 32 informantes, levando em conta os seguintes aspectos: a) sexo - para cada ponto, foram entrevistados dois homens e duas mulheres (um homem e uma mulher de cada geração), totalizando 4 informantes por localidade; b) faixa etária - foi distribuída em duas gerações: jovens de 18 a 32 anos (G1) e adultos de 48 a 62 anos (G2). Em cada ponto, foram selecionados dois informantes da G1 e dois da G2; c) escolaridade - foram escolhidos os informantes com escolaridade igual ou inferior ao 9º ano do ensino fundamental.

Para todos os informantes, foram aplicados dois tipos de questionários, um questionário fonético e um semântico e lexical, com o objetivo de coletar dados, possibilitando, assim, a elaboração das 147 cartas linguísticas sendo que 84 léxicas e 63 fonéticas. As cartas fonéticas foram utilizadas para extrairmos os fenômenos fonéticos presente na variedade linguística potiguar.

Além desse atlas, utilizamos também o Atlas Geolinguístico do Litoral Potiguar que, segundo Pereira (2008)[7], apresenta dados para estudos da variedade linguística na forma da coleta de dados e no tratamento cartográfico de variantes faladas no litoral leste do Rio Grande do Norte. Nele podemos ver a contribuição no registro, na análise e na descrição das variantes linguísticas fonético-fonológicas, léxico-semânticas e morfossintáticas, tanto no ponto de vista diatópico e quanto no diastrático.

Essa autora acrescenta que o estabelecimento da rede de pontos, sugerida por Antenor Nascentes para os Atlas Linguísticos, foi feito segundo alguns critérios, tais como: (a) densidade demográfica, (b) a história do município, (c) aspectos geográficos e (d) importância econômica da localidade que representa o universo de pesquisa (ou pontos de pesquisa).

7 Neste subtópico, apresentamos as principais informações expostas por Pereira (2008). Para maiores informações sobre a pesquisa feita, a tese da autora deve ser consultada.

Figura 3 – Mapa da região do litoral potiguar.

Fonte: Pereira (2008, p. 16).

Considerando esses critérios, foi investigada a fala de 24 informantes em cinco municípios, dentre os 11 (onze) que constituem a rede de pontos do Projeto Atlas Linguístico do Rio Grande do Norte, pertencentes a uma área chamada convencionalmente de litoral potiguar. A Figura 3 mostra a área que compreende os municípios onde foi coletada a pronúncia da fala potiguar e relaciona os números a cidades dessa mesorregião. A capital Natal corresponde ao número 1, Canguaretama ao número 2, Touros ao número 3, Macau ao número 4 e Areia Branca ao número 5.

Para Pereira (2008), os resultados do estudo foram obtidos por meio de entrevistas *in loco* e posterior análise dos fatores que condicionam os possíveis traços de variação em níveis fonético, léxico e morfossintático que se pretendiam detectar. Os informantes foram selecionados com base em alguns critérios adotados pelo Projeto ALiB e a coleta foi feita através dos questionários utilizados na recolha dos dados para a construção do *corpus*. O Questionário Fonético Fonológico (QFF) é composto de 35 perguntas, o Questionário Semântico Lexical (QSL) é constituído de 35 perguntas e o Questionário Morfossintático (QMS) é formado por 10 questões. Desses questionários, damos ênfase apenas ao primeiro, uma vez que nele estão presentes as formas fonéticas transcritas da fala potiguar.

Esses atlas linguísticos apresentam uma visão geral da fala potiguar de forma consistente e os dados serviram para a identificação e para a quantificação dos fenômenos

linguísticos, uma vez que podemos encontrar a transcrição das palavras nos símbolos do AFI, que correspondem à produtividade fonética pelo informante, localidade e os gráficos com base em índices percentuais de ocorrências.

4. Regras da fonologia gerativa

Para a implementação do sistema Potigrafone, fundamentamos na fonologia gerativa de Chomsky e Halle (1968) e em Schane (1975). Para Chomsky e Halle (1968), o componente fonológico é um sistema de regras que relaciona estruturas de superfície a representações fonéticas. Nesse caso, as regras constituem a gramática de um determinado falante da língua e determinam em detalhes a forma das frases com que o falante produzirá e compreenderá a estrutura linguística. Uma forma direta de efetuar as mudanças em uma gramática é adicionar novas regras ao componente fonológico, atuando na modificação, no acréscimo ou no apagamento dos segmentos das formas de entrada.

Nos subtópicos a seguir, enfatizamos a formalização das regras fonológicas para o gerativismo e apresentamos a execução do sistema partindo das representações subjacentes – ou *input* – para as representações de superfície – ou *output*.

4.1 Regras fonológicas

Schane (1975) apresenta as três regras da fonologia gerativa: a regra de transformação, de inserção e de cancelamento dos segmentos. Ressaltamos que essas regras foram desenvolvidas e detalhadas no capítulo três por Chomsky e Halle (1968). O quadro 3 abaixo apresenta os principais símbolos utilizados na formalização dessas regras fonológicas.

Quadro 3 – Símbolos utilizados nas regras fonológicas.

Símbolo	Nome	Descrição
_	traço	Serve para marcar a posição exata em que ocorre o Segmento, cujo contexto será caracterizado pelo que o precede e segue.
.	Ponto	Separador silábico
+	Kleene plus	Marca a fronteira entre as sílabas.
/	Barra inclinada à direita	Serve para separar o ambiente do segmento fonônico do restante da regra.
Ø	Conjunto vazio	A eliminação é indicada pelo Ø, mas poderiam ser utilizados *colchetes vazios* ou 0 (*zero*)
$	cifrão	Representa a sílaba.

→	seta	Substituição direta em contexto significa que o elemento à esquerda transforma-se no da direita.
#	Cerquinha	Limite de palavra

Fonte: adaptado de Cagliari (2002, p.29)

a) regra de transformação

/l/ → [w] / _.$

Lê-se essa regra da seguinte forma: o segmento /l/ transforma-se em glide [w], quando estiver em posição final de sílaba. Exemplo: /sal/ → ['saw] e /salta/ → [saw'ta].

b) regra de cancelamento

/l/ → Ø / 'V_+s

Lê-se essa regra da seguinte forma: o segmento /l/ é cancelado quando precedido de vogal acentuada e seguido do morfema de plural s. Essa regra não é válida para as palavras adjetivas terminadas em <vel>, como <amável>. Porém, na palavra <sal> pode ser aplicada. Exemplo: 'sal+S → ['sa+S].

c) regra de inserção

Ø → [i] / 'V_.+s

Lê-se essa regra da seguinte forma: insira o segmento [i] quando uma vogal acentuada é seguida do morfema de plural S. Exemplo: 'sa+S → ['saiS] “sais”.

O uso dessas regras tem o propósito de explicitar o mecanismo fonológico e fornecer as descrições estruturais determinantes das mudanças que ocorrem entre as representações gráficas e as representações fonéticas da fala potiguar. Essas regras não são aplicadas ao acaso, mas segundo critérios que satisfaçam determinadas exigências. Além disso, as "regras fonológicas expressam processos fonológicos e idealmente o fazem de maneira simples, econômica e em caráter generalizador." (SILVA, 2014, p. 198).

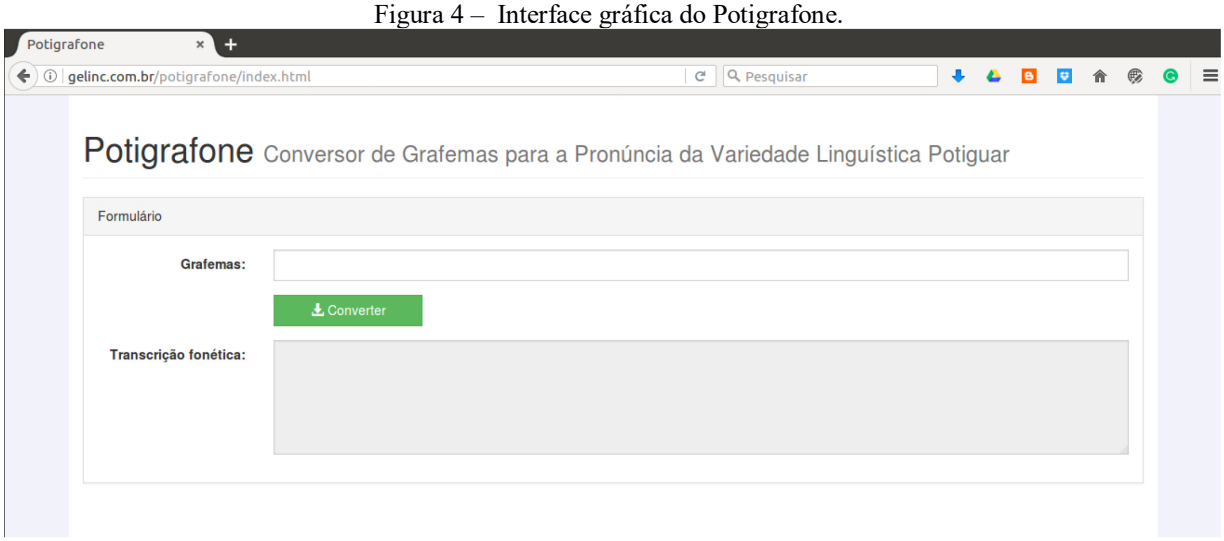
No próximo subtópico, veremos como essas regras foram aplicadas ao sistema para a conversão de grafema para a representação fonética da variedade linguística potiguar.

4.2 Representações subjacentes e representações de superfície

O *Potigrafone* relaciona os símbolos de entrada com os símbolos fonéticos do SAMPA aplicando as regras fonológicas e retorna a transcrição fonética para a variante linguística potiguar. Esse sistema é executado numa interface gráfica na linguagem de programação html 5 e está disponível na web por meio do endereço eletrônico: http://gelinc.com.br/potigrafone/index.html.

O uso do sistema é muito intuitivo e tem apenas dois campos: um para a entrada da palavra escrita em português e outro para a saída da transcrição fonética, como podemos ver na figura 3. No entanto, para melhor compreensão da execução do sistema e da aplicação das regras fonológicas, utilizaremos o exemplo da palavra <caixa>, apresentada no quadro 3. Essa palavra apresenta duas variantes fonéticas na fala potiguar: ["kajS6] e ["kaS6].

Figura 4 – Interface gráfica do Potigrafone.



Fonte: elaborada pelo próprio autor.

O *input* <caixa>, escrito no campo da entrada de dados "grafemas", é relacionado aos elementos do alfabeto da língua portuguesa existentes no arquivo de entrada e o sistema reconhece cada letra da palavra como pertencente ao alfabeto da língua formal. "Se um dos símbolos de entrada não pertencer ao símbolo de entrada, o sistema responde ao usuário com o conjunto de caracteres "???", indicando que a palavra é estranha à língua formal.

Após o reconhecimento dos caracteres, o sistema executa a separação silábica da palavra, inserindo o símbolo de igualdade "=" entre as sílabas. Esse sinal não aparece na transcrição fonética, todavia, o Potigrafone utiliza-o para a aplicação das regras, ou seja, ele serve como elemento de fronteira silábica.

Feita a divisão silábica, o conversor acessa o conjunto de regras fonológicas e aplica primeiramente as regras de acentuação tônica e depois as regras de correspondência ao caractere de entrada. Para essa situação, o sistema faz a acentuação tônica das palavras atribuindo à sílaba tônica o diacrítico do SAMPA, a aspa dupla " " ". Com a marca da tonicidade, a palavra apresenta a seguinte forma: <"cai=xa>, pois temos a separação silábica e marcação tônica.

Feita inserção desses dois sinais diacríticos, o sistema passa à execução das regras de transformação e dos fenômenos fonéticos. A letra <c> é transformada no som oclusivo [k], pois o conversor atribui a essa consoante a regra da consoante oclusiva que está expressa:

c -> k || _[vpos | vmed | liq | "="];

onde se lê: a letra <c> será transformada em oclusiva velar desvozeada [k], quando estiver antecedida das vogais médias e posteriores ou das consoantes líquidas ou do sinal de separação silábica. Desse modo, essa regra é válida para esse contexto e também para outros, como nas palavras <co=po>, <cri=a>, <cac=to>.

Depois da transformação da oclusiva, o conversor aplica a regra do núcleo silábico ao par de vogais <ai>. Nesse contexto, a letra <a> não recebe nenhuma transformação, porque ela é núcleo silábico do ditongo, mas a letra <i> é transformada em um glide [j], como explicita a seguinte regra:

i -> j || [vogal]_ ;

onde se lê: a letra <i> será transformada em glide [j], quando estiver seguida de uma vogal. Logo, a transcrição para esse ditongo é a vogal baixa [a] seguida de glide [j], [aj].

Após a transformação da vogal alta anterior em glide, o Potigrafone faz a transformação do grafema <x> em uma fricativa alveolar desvozeada [S], aplicando a seguinte regra

x -> S || [.#. | .#. """ | [vogal vogal] "="]_;

onde se lê: a letra <x> será transformada em consoante fricativa, quando estiver antecedida por uma fronteira de palavra, ou seja, no início de uma palavra, ou pelo início de uma palavra seguida de um sinal de tonicidade, ou pelo ditongo seguido de um separador silábico. Assim, essa regra aplica-se a outros contextos para esse grafema, como na palavra <xícara>.

Por último, o sistema executa a transformação do grafema <a> em uma vogal média átona, apresentada pelo símbolo <6> do SAMPA. Para isso, o Potigrafone executa a regra seguinte, que mostra a redução vocálica para essa vogal.

a->6 || _(s) .#. ;

onde se lê: a vogal <a> será reduzida quando estiver numa sílaba pós-tônica final antecedida ou não do grafema <s>.

Apenas com a utilização dessas regras, a saída do sistema ficaria da seguinte forma: ["kaj=S6]. Nesse caso, teríamos uma saída para a transcrição da palavra <caixa> e com a separação silábica. No entanto, na fala potiguar, temos outra variante para essa palavra, a pronúncia com a forma monotongada. Para a aplicação do fenômeno da monotongação, o Potigrafone executa a seguinte regra de substituição:

```
j->[ j | 0] || _ "=";
```

onde se lê: a semivogal [j] receberá duas saídas: a semivogal [j] e a ausência de representação sonora [0], quando estiver precedida pelo sinal de separação silábica. Desse modo, a saída do sistema ficaria da seguinte forma: ["kaj=S6], ["ka=S6].

Feito isso, o Potigrafone executa a última regra que é a de cancelamento do sinal de separação silábica,

```
["="] -> 0;
```

onde se lê: os sinais de separação silábica devem ser cancelados. Assim, finaliza o processo de transcrição fonética automática da palavra <caixa>, retornando ao usuário duas formas fonéticas da variedade linguística potiguar: ["kajS6], ["kaS6].

5. Considerações finais

Neste trabalho, apresentamos um conversor de transcrição fonética automática para a variedade linguística potiguar e apontamos as seguintes considerações. A primeira consideração que fazemos diz respeito à diferença na transcrição fonética entre Potigrafone e o Petrus. As formas fonéticas do primeiro sistema são constituídas por símbolos e diacríticos do alfabeto fonético SAMPA e podem apresentar mais de uma variante linguística potiguar para uma mesma forma grafema; enquanto que o segundo apresenta os símbolos do Alfabeto Fonético Internacional e, com exceção do fenômeno da alternância vocálica, mostra apenas uma variante paulista, mas executa outras funções como a separação silábica e a classificação da palavra.

A segunda se relaciona à apresentação dos aspectos gráficos do sistema. Aqui, apresentamos apenas algumas particularidades das consoantes no desenvolvimento de um conversor de transcrição fonética e deixamos de lado os aspectos vocálicos que são mais ricos e mais complexos. No entanto, recomendamos a leitura da tese de Carvalho (2016) para maior conhecimento sobre a forma de implementação e de apresentação das vogais nesse conversor.

Em terceiro lugar, consideramos que, diferentemente de outros conversores, o Potigrafone foi desenvolvido a partir de um *corpus* da variedade linguística para conhecer as variantes faladas no estado. Assim, a transcrição espelha a fala potiguar. Por fim, esse conversor está disponível para que os usuários possam utilizá-lo como auxílio também em suas pesquisas.

Referências Bibliográficas

BEESLEY, K. R.; KARTTUNEN, L. **Finite-State Morphology**: Xerox Tools and Techniques, 2002.

CAGLIARI, L. C. **Análise fonológica***:* introdução à teoria e à prática, com especial destaque para o modelo fonêmico. Campinas-SP: Mercado de Letras, 2002.

CARVALHO, C. I. C. **Transdutor de estados finitos para conversão de grafema para a pronúncia da variedade linguística potiguar**. 2016. 160 f. Tese (doutorado em Linguística) – Universidade Federal do Ceará, Centro de Humanidades, Departamento de Letras Vernáculas, Fortaleza.

CHOMSKY, N.; HALLE, M. **The sound pattern of english**. New York: Harper e Row, 1968.

MARQUIAFÁVEL, V.; BOKAN, A.; ZAVAGLIA, C. PETRUS: A rule-based grapheme-to-phone converter for Brazilian Portuguese. In: **Workshop on Computational Processing of the Portuguese Language** – PROPOR, 11, 2014, São Carlos-SP. Disponível em: http://www.lbd.dcc.ufmg.br/colecoes/propor-demo/2014/004.pdf. Acesso: 09 de agosto de 2015.

MARQUIAFÁVEL, V.; ZAVAGLIA, C. Transcrição fonética automática para lemas em verbetes de dicionários do Português do Brasil. in: **Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology**, Cuiaba - MT, Brasil, outubro 2011, p. 154-158. disponível em: http://nilc.icmc.sc.usp.br/til/stil2011_English/stil/artigos/Short/STIL2011_SP1.pdf. Acesso em: 10 de agosto de 2015.

PEREIRA, M. das N. **Atlas geolinguístico do litoral potiguar - ALiPTG**. 2008. 136 f. volumes I e II, Tese (doutorado em Linguística) - Universidade Federal do Rio de Janeiro, Faculdade de Letras, Rio de Janeiro, 2008.

SCHANE, S. A. **Fonologia gerativa**. Trad. Alzira S. da Rocha, Helena M. Camacho, Junéia Mallas. Rio de Janeiro: Zahar Editores, 1975.

SEARA, I. C.; NUNES, V. G.; LAZZAROTTO-VOLCÃO, C. **Para conhecer fonética e fonologia do português brasileiro**. São Paulo: Contexto, 2015.

SILVA, M. B. da . **Atlas linguístico do centro-oeste potiguar**. 2012. 327 f. Tese (doutorado em Linguística). Centro de Humanidades, Universidade Federal do Ceará, Fortaleza, 2012.

SILVA, T. C. **Fonética e fonologia do português**. 10. ed. São Paulo: Contexto, 2014.

TÓS, M. L. R. D. Monotongação: estudos da variação dos ditongos nos falares do interior paulista nos dados do Alib. In: **Seminário do GEL**, 58., 2010, São Carlos (SP): GEL, 2010. Disponível em: http://www.gel.org.br/?resumo=6413-10. Acesso em: 26.09.2016.

WELLS, J. C. **Computer-coding the IPA**: a proposed extension of SAMPA. University College London. disponível em: https://www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf. Acesso em: 30 agosto 2014.

Artigo recebido em: 14.01.2017 Artigo aprovado em: 05.04.2017