VERA VASILÉVSKI

CONSTRUÇÃO DE UM SISTEMA COMPUTACIONAL PARA SUPORTE À PESQUISA EM FONOLOGIA DO PORTUGUÊS DO BRASIL

FLORIANÓPOLIS
2008

VERA VASILÉVSKI

CONSTRUÇÃO DE UM SISTEMA COMPUTACIONAL PARA SUPORTE À PESQUISA EM FONOLOGIA DO PORTUGUÊS DO BRASIL

Tese apresentada ao Curso de Pós-graduação em Lingüística da Universidade Federal de Santa Catarina (UFSC), como requisito para obtenção do título de Doutor em Lingüística; área de concentração: Lingüística aplicada; linha de pesquisa: *Corpus* e gênero: análise e aplicações.

Orientador: Prof. Dr. Marco Antônio Esteves da Rocha

Co-orientadora: Profa. Dra. Leonor Scliar-Cabral

FLORIANÓPOLIS
2008

VERA VASILÉVSKI

CONSTRUÇÃO DE UM SISTEMA COMPUTACIONAL PARA SUPORTE À PESQUISA EM FONOLOGIA DO PORTUGUÊS DO BRASIL

Esta tese foi aprovada como requisito para a obtenção do título de Doutor em Lingüística, pelo programa de Pós-graduação em Lingüística da Universidade Federal de Santa Catarina.

Prof. Dr. Fábio Lopes da Silva
Coordenador do Programa de Pós-graduação em Lingüística
da Universidade Federal de Santa Catarina

Banca examinadora:

Prof. Dr. Marco Antônio Esteves da Rocha
Orientador
Universidade Federal de Santa Catarina

Profa. Dra. Leonor Scliar-Cabral
Co-orientadora
Universidade Federal de Santa Catarina

Prof. Dr. Fábio Lopes da Silva
Universidade Federal de Santa Catarina

Profa. Dra. Regina Lamprecht
Pontifícia Universidade Católica do Rio Grande do Sul

Profa. Dra. Izabel Seara
Universidade Federal de Santa Catarina

Prof. Dr. André Berri
Universidade Federal de Santa Catarina

Florianópolis, 16 de junho de 2008.

Aminhas irmãs Márcia de Lima e Mariza Karine de Lima,

as melhores amigas que uma garota pode ter,

porque sempre seguram barras,

e nunca seguram lágrimas.

Amo vocês.

AGRADECIMENTOS

A Deus, maior energia positiva do universo, agradeço a fé que me fez continuar nessa empreitada que, por várias vezes, achei que não concluiria.

A minha mãe, Lúcia Vasilévski, a força de vontade e a firmeza, que não é mole, não!

A meu irmão, Marcelo Florêncio, meu eterno nenê, ser puro *rock'n roll*, futebol (dá-lhe Furacão!) e tranqüilidade, para me alegrar a qualquer momento.

A meu pai, João de Lima, agradeço o carinho e o amparo em todas as horas, desde o começo.

A meus amigos florianopolitanos, Manoel de Jesus, Cecília Vieira e Regina Gerber, a amizade, que é uma conquista do ser humano e mostra que ele tem algo de bom, que faz bem para outras pessoas, e isso dá fôlego para prosseguir.

A minha alegre e querida amiga do Ceará, que pousou na UFSC, Glória Amaral, agradeço o apoio maternal dado no início do doutorado, que tanto facilitou a trajetória.

A minha linda e querida amiga de São José do Cedro, Silene Etges, agradeço tanto carinho e apoio fraternal que ganhei no final dessa jornada.

À festiva amiga curitibana Sandra Mello agradeço os cuidados de irmã mais velha, os cafés, as danças e os vinhos animados, regados a conversas e risadas.

À amiguinha curitibana Ingrid Mello, a ajuda na feitura do logotipo do sistema Nhenhém e as cantorias.

Aos pequerruchos amigos curitibanos Heitor Ruggeri e Helena Ruggeri, a companhia e os divertidos passeios, que quebraram a monotonia.

Aos sobrinhos Christian Lima, Dudu Portugal e Maikon Portugal (meu afilhado) o carinho e entenderem minha ausência em muitos momentos.

A Ronaldo Ribeiro, o passado frutuoso que marcou a abertura da clareira.

Às amigas catarinenses de graduação Saionara Gréggio e Carina Merkle, as conquistas e os conhecimentos compartilhados, que se fizeram importantes no doutorado, depois de tanto tempo.

A curitibana Simone Doubek, minha primeira chefe e amiga para sempre, me ensinar a ter responsabilidade e a fazer tudo bem feito.

Ao Márcio Araújo, programador em Delphi, a ajuda técnica e o entusiasmo com a língua portuguesa, que falta a muitos pesquisadores e professores da área. Conhecê-lo nessa caminhada foi tirar a sorte grande. Valeu, rapaz, você é bom!

A meu orientador, Marco Rocha, a liberdade para arriscar alto por idéias lingüísticas incertas, as não-imposições e o respeito por meu jeito de escrever.

A minha co-orientadora, Leonor Scliar-Cabral, agradeço a dedicação, o apoio e o entusiasmo contagiante.

Aos professores André Berri, Regina Lamprecht e Izabel Seara, terem aceitado o convite para compor a banca.

Aos professores Fábio Lopes e Kanavinil Rajagopalan, o crédito e o respeito dados a minhas opiniões "absurdas" e terem me dado voz e vez.

Ao professor Luiz Carlos Cagliari agradeço a atenção a mim dispensada na resolução de dúvidas fonético-fonológicas insistentes e o incentivo ao trabalho com fonologia.

Aos professores de alfabetização e amigos que testaram o Nhenhém e deram sua opinião, a boa vontade e a animação.

...

*V*ozes veladas

veludosas vozes

volúpias dos violões

vozes veladas

vagam

nos velhos vórtices velozes

dos ventos

vivas

vãs

vulcanizadas

...

Cruz e Sousa

RESUMO

Esta pesquisa apresenta e discute a construção de um programa computacional que descodifica o sistema verbal escrito oficial do português do Brasil em símbolos fonológicos. A motivação para construir o programa veio do alto grau de transparência do sistema alfabético do português do Brasil, que levou à suposição de que tal transparência corresponde ao nível em que o sistema alfabético é lógico. Assim, a hipótese que norteia a pesquisa é a de que o alto nível de previsibilidade dos valores que os grafemas do sistema alfabético do português do Brasil têm pode ser reproduzido em um sistema computacional baseado em regras, que faz a conversão grafema-grafofonema de modo automático. Metodologicamente, o desenvolvimento do aplicativo associa lingüística computacional, lingüística de *corpus*, estatística, fonética e fonologia. Por ser planejado com base sólida em metodologia e teoria lingüística apropriadas, o aplicativo pôde ser construído em linguagem de programação de computadores não específica para tratar a língua. A conversão baseia-se em um alfabeto fonológico, em que se usam as fontes do International Phonetic Alphabet (IPA). O aplicativo lê massas de dados relativamente grandes e fornece relatórios de conversão fonológica e relatório estatístico dos fonemas dos textos convertidos. Juntamente com o programa, dão-se alguns passos para se desenvolver metodologia própria para formação e tratamento de *corpus* lingüístico fonológico. Testes preliminares, feitos a partir de um *corpus* fonológico montado com base nos pressupostos da teoria pertinente, indicam que o aplicativo reproduz a parte do sistema verbal escrito que é previsível por regras de descodificação fonológicas, com aproximadamente 98% de acerto, e o sistema verbal escrito todo, com acerto mínimo de 95%. Ainda, o programa fornece a tonicidade das palavras da língua portuguesa com acerto superior a 99%, e o relatório estatístico mostra os padrões de distribuição fonêmica do português. A pesquisa revela que esses porcentuais aumentarão mediante novos estudos, portanto, comprova-se a hipótese inicial e confirma-se que a elevada previsibilidade do sistema alfabético do português do Brasil dá-se por ele ser fundamentado em intuições fonológicas. Um aplicativo dessa natureza pode ser ferramenta auxiliar na pesquisa e no ensino de alfabetização, na fonologia, na literatura e em outras áreas.

Palavras-chave: Sistema alfabético do português do Brasil; lingüística computacional; lingüística de *corpus*; intuições fonológicas; previsibilidade.

ABSTRACT

This research presents and discusses the construction of a computational program that decodes Brazilian's official writing system into phonological symbols. What inspired the software development was the high level of transparency of Brazilian Portuguese alphabetic system, which led to suppose that such transparency corresponds to the level in which this system is based on Logics. So, the research investigates the hypothesis that the high level of predictability of the values that the Brazilian Portuguese alphabetic system graphemes bear can be reproduced by a software based on rules, that makes the conversion of graphemes into phonemes automatically. Methodologically, the applicative development associates Computational Linguistics, Corpus Linguistics, Statistics, Phonetics, and Phonology. Since the program planning combined proper methodology and linguistic theory, the software could be built in a computer programming language which is not specifically planned for the treatment of human language. The program bases the translation on a phonologic alphabet, which takes into account the International Phonetic Alphabet (IPA) fonts. The software reads relatively huge bunches of data, and bestow phonologic reports, and statistical reports. Some directions in order to develop specific methodology to form a phonologic corpus and deal with it are given. At examining a phonologic corpus rightly assembled, preliminary tests done by drawing on the applicative showed that it reaches approximately 98% of accuracy, at reproducing the portion of the Brazilian writing system that is predictable by decoding rules, and in relation to writing system as a hole the correctness is not less than 95%. Also, the program reaches 99% of precision at putting words accent. Some new studies tend to increase these numbers. The statistical report displays Portuguese language phoneme patterns of distribution. In conclusion, the research confirms the hypotheses, and authenticates that the high level of predictability of Brazilian alphabetic system is due to it be based on phonologic intuitions. A software like this can be a tool for aiding researchers and teachers who work with literacy, Literature, Phonology, Phonetics, among other areas.

Keywords: Brazilian Portuguese alphabetic system; Computational Linguistics; Corpus Linguistics; phonologic intuitions; predictability.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE QUADROS E TABELAS

INTRODUÇÃO

Este estudo apresenta e discute a construção de um programa eletrônico que descodifica o sistema verbal escrito oficial do português do Brasil em símbolos fonológicos, de acordo com o que é previsível na língua portuguesa. Trata-se do sistema eletrônico de conversão grafema-fonema Nhenhém, em sua versão 1.0.

Esse programa é resultado do projeto que originou esta tese, o qual se intitula *Estudo exploratório de características fonêmicas do português a partir de processamento computacional*. A proposta era desenvolver um programa de computador para sistematizar características fonêmicas da língua portuguesa com base em massas de dados lingüísticos. O aplicativo deveria descodificar textos segundo as características fonológicas deles – ou seja, converter automaticamente a língua escrita em símbolos fonológicos – e permitir manipulação dos textos convertidos, operações de armazenamento deles em disco e gerar relatórios de caracteres e de estatística dos traços distintivos dos fonemas que compõem esses textos.

Envolvem-se lingüística computacional, lingüística de *corpus* e fonologia em uma única pesquisa. A idéia de realizar um trabalho que associasse essas três áreas surgiu do contato da pesquisadora com elas, nas quais realizou alguns estudos. Inicialmente, houve o conhecimento e a aplicação do inseparável par lingüística computacional e lingüística de *corpus*, que norteou sua dissertação de mestrado, na qual a lingüística computacional foi restrita, mas o uso de *corpus* foi enfatizado, discutido e aplicado. Após isso, outros estudos teóricos (VASILÉVSKI, 2007a e 2007b) e práticos (VASILÉVSKI, 2007c) nessa área foram desenvolvidos. Estudos em fonologia vieram depois e, imediatamente, com seu aprendizado, ficou claro que seria frutífera a junção das três áreas para pesquisar a língua, posto que a fonologia, dadas suas delimitações, apresentou-se promissora para ser investigada por meio de um programa eletrônico.

Houve mais motivação para investir nesse projeto. A época atual, a era da informação, com as tecnologias de computação e de redes, com transmissão de dados,

informações em tempo real, linguagens de programação acessíveis, não poderia ser mais propícia para incentivar o desenvolvimento de um programa lingüístico eletrônico como o Nhenhém. Dicionários eletrônicos em programas e disponíveis para consulta na Rede possibilitaram pesquisar listas de palavras classificadas e foram ferramenta para conferência de dados. A Internet é o maior banco de informações do mundo, e foi fonte de conhecimento indispensável em todas as áreas que esta pesquisa envolve – suporte técnico-científico 24h –, então, mais motivação.

No entanto, foi decisivo para idealizar o programa tomar conhecimento da clareza do sistema ortográfico do português, tendo em vista seus princípios, que revelam intuições fonológicas. O nível de transparência de um sistema ortográfico mensura-se pela previsibilidade dos valores de seus grafemas. Nesse sentido, o sistema alfabético do português do Brasil tem grande transparência.

Daí se tem que, se o nível de previsibilidade do sistema alfabético do português do Brasil é elevado, então, esse sistema é lógico, ou seja, fundamenta-se em regras sólidas, em grande medida. A partir disso, emerge a hipótese de que é possível criar um programa de computador baseado em regras que reproduza o sistema ortográfico do português do Brasil e obtenha, na tradução eletrônica grafema-fonema, acerto proporcional a tal nível de previsibilidade.

O maior obstáculo para realizar um aplicativo como esse são os conhecimentos exigidos, pois há reconhecida distância entre os lingüistas e a computação, bem como os programadores não têm conhecimento de lingüística, mas geralmente encabeçam os raros projetos de desenvolvimento de sistemas dessa natureza. Contribuir no sentido de transpor essa barreira é outro estímulo a esta proposta, tendo em vista que, fora do campo acadêmico, a pesquisadora já desenvolveu sistemas que gerenciam bancos de dados, em Visual Basic, e conhece outras linguagens de programação de computador.

Ainda, o desejo de unir conhecimentos acumulados na área tecnológica com os da área de ciências humanas impulsionou o projeto, afinal, parodiando Lavoisier, assim como na natureza, na vida, o aprendizado nunca se perde, transforma-se. Conhecimentos em áreas distintas foram reunidos para compor esta tese: física, matemática, computação, língua inglesa, língua portuguesa, fonética, fonologia. Da

bagagem de conhecimento da pesquisadora, tudo contribuiu para a realização da tese: o curso de Técnico em Eletrotécnica (CEFET-PR), a experiência como eletricista, como instrutora de informática, como professora de português e de inglês, como desenvolvedora de banco de dados, o curso de Letras, o mestrado, as disciplinas extracurriculares de gramática e as disciplinas do doutorado, obviamente.

A falta de apoio de pesquisadores da área de fonética e fonologia, sobretudo dos que trabalham com sistemas eletrônicos, seria um desestímulo, que se dissolveu, mediante o incentivo e a empolgação de outros pesquisadores.

Com o projeto em andamento, à medida que os estudos se desenvolviam, observava-se que não havia, pelo menos não era acessível, metodologia científica para tratar e armazenar textos fonológicos em massa. Isso foi previsto no projeto de tese. Então, a metodologia para trabalho com *corpus* para pesquisa em fonologia foi desenvolvida para este estudo, e é parte de seu conteúdo. Por causa disso e das noções de linguagem de programação de computadores, a teoria confunde-se com a metodologia e vice-versa, algumas vezes. Nesse ínterim, também ficou clara a importância de saber acerca da transformação do sistema verbal escrito através do tempo, para que se rastreassem e discutissem algumas questões lingüísticas notáveis.

Estudos e testes preliminares mostraram que seria necessário, além de bibliografia específica sobre fonética e fonologia, apoio da filologia, lexicografia, Nomenclatura Gramatical Brasileira e, conseqüentemente, gramática normativa, porque um sistema eletrônico trabalha com lógica, e a lógica é objetiva, portanto, precisa de padrões e normas taxativos, para que se reduzam inconsistências. Definições abertas, sensíveis e amplas não podem ser transformadas em lógica. Normas categóricas tornam o sistema mais consistente, pois o computador resume tudo a 0 e 1. Algumas definições que podem receber críticas tiveram de ser tomadas, mas sempre com respaldo em alguma teoria e em prol da eficiência do programa. Não se prefere uma área, teoria ou autor em detrimento de outros, extrai-se de cada um o que é relevante aos propósitos da pesquisa. O compromisso é com a ciência e com a sociedade.

Assim sendo, a exposição de contratempos que surgiram na feitura do programa e de alguns princípios de programação de computador objetiva divulgar informação sobre o funcionamento da programação, sobre a forma como a máquina processa a língua. A partir disso, compreendem-se um pouco as dificuldades enfrentadas pelos programadores, que têm de entender o que o pesquisador ou cliente quer e planejá-lo, convertê-lo mentalmente para a forma como a máquina pode fornecer a resposta desejada.

Os empecilhos aumentam sobremaneira, se o pesquisador não tem noção de programação de computadores, porque ele não consegue se fazer entender pelo programador, que dificilmente tem conhecimentos de lingüística. Então, dentre as dificuldades possíveis para o desenvolvimento de um programa eletrônico que envolva lingüística, está a comunicação entre as partes envolvidas: lingüista-programador-máquina. O elo entre eles tem de ser forte para asseverar êxito na pesquisa interáreas.

Depois de dois anos e meio de estudos em lingüística e programação, testes no papel, tentativas e erros, chegou-se a um projeto de sistematização das principais regras de descodificação do português do Brasil em ambiente eletrônico. Depois dessa parte teórica, passou-se à prática, ou seja, testes em computador, na qual o projeto teve de ser adaptado em vários aspectos, tendo em vista incompatibilidades entre o ambiente físico e o virtual, entre as configurações do computador e a linguagem de programação, entre ambos e a fonte de fonética em que a tradução seria feita. Levou mais seis meses para esses ajustes serem concluídos.

Então, em um ano, em que houve muitas idas, voltas, revoltas, rodeios e reestruturações, o trabalho foi com a linguagem de programação escolhida para se desenvolver o sistema eletrônico de conversão grafema-fonema. Quando o programa chegou ao nível de descodificação de em torno de 50% do texto, ele mereceu ser batizado, enquanto se trabalhava no aumento desse porcentual. Assim foi, até que o sistema de conversão grafema-fonema, então, Nhenhém©, chegou à versão 1.0. O nome dado ao programa, “nhenhém”, vem do tupi e significa o prosseguir, o repetir-se interminavelmente de um ato – como expressar-se –, um movimento – como o dos lábios –, um som – como a voz –, portanto, “falar, falar”.

A relevância desta pesquisa reside: na associação proposta entre lingüística de *corpus*, fonologia e computação, que beneficia os estudos da língua; no estudo teórico realizado sobre a aplicação da lingüística de *corpus* ao trabalho com fonologia; na experiência de uso do alfabeto fonético na estatística; e na aplicação dos dados dos relatórios do programa em pesquisas lingüísticas. A partir da análise de relatórios do programa, pode-se estimar em porcentual o nível de transparência do sistema alfabético do português do Brasil. O aplicativo será disponibilizado para pesquisadores e professores que trabalham com alfabetização, que poderão extrair informações dos dados de trabalho de forma inédita.

ORGANIZAÇÃO DO ESTUDO

Esta tese divide-se em cinco capítulos principais. A primeira seção expõe a revisão da literatura com aspectos históricos acerca da língua, que serviram de base para a organização teórica do programa Nhenhém e conscientização sobre o tema da pesquisa. A segunda seção é a metodologia, em que se documentam decisões tomadas no intuito de sistematizar cientificamente em ambiente eletrônico as regras de descodificação que regem a língua portuguesa escrita do Brasil, bem como a montagem de massa lingüística fonológica e aplicação de estatística a ela. A terceira seção traz a descrição funcional do programa desenvolvido. O quarto capítulo apresenta discussões e reflexões sobre a construção do Nhenhém e resultados práticos. A conclusão aborda questões lingüísticas polêmicas relacionadas ao comportamento do programa e expõe sugestões de complementos futuros. Finalizam o texto as considerações finais.

1 ASPECTOS HISTÓRICO-TEÓRICOS

as coisas estão pretas
uma chuva de estrelas
deixa no papel
esta poça de letras

Leminski (2008)

Apresenta-se uma revisão da literatura relacionada a aspectos da língua portuguesa relevantes para a pesquisa, como sua história, sistema escrito oficial, língua falada, fonologia, controvérsias lingüísticas, definições lingüísticas, dentre outros. Também se trata de teoria computacional relacionada ao sistema eletrônico desenvolvido. É a base para compreensão dos princípios e do uso do Nhenhém 1.0.

Adotam-se algumas convenções para distinção entre escrita e fala. A forma gráfica de palavras e morfemas é indicada por aspas inglesas ("festa"). Palavras e morfemas não pertencentes ao português estão em itálico (*byte*), a menos que sejam nomes próprios. Letra e grafema referem-se à língua escrita, enquanto fonema refere-se à língua falada. As referências a ditongo, tritongo, hiato, vogal e consoante podem referir-se tanto à modalidade escrita como falada da língua, o contexto esclarecerá de qual se trata. Semivogal é referência fonêmica. Na escrita, as semivogais equiparam-se às vogais "i" e "u", ou à consoante "l", ou ainda às consoantes "m" e "n.

1.1 HISTÓRIA COMPACTA DA LÍNGUA PORTUGUESA EM PORTUGAL E NO BRASIL[1]

A história de uma língua não é um esquema rigorosamente preestabelecido, não é um problema algébrico. Não se pode partir do latim e chegar diretamente aos dias de hoje, saltando por vários séculos de intensa vida. A evolução é complexa e melindrosa,

[1] As informações contidas nesta seção apóiam-se em Silva Neto (1988), Câmara Jr. (1986), Carvalho (1996), Scliar-Cabral (2003), Cruz (2007), jornal *O Globo* (2008 e 2007), Wikipédia (2007), Brasil (2007) e Malha (2007).

relaciona-se a muitos acidentes, porque não representa a evolução de algo pronto e acabado, mas sim as vicissitudes de uma atividade em perpétuo movimento.

Alguns fatos históricos contribuíram para a formação da língua portuguesa: a conquista romana da Península Ibérica, a invasão dos bárbaros germanos, a constituição dos impérios bárbaros, como o visigótico, o domínio árabe na península, a luta da reconquista cristã, a formação do reino de Portugal e a expansão ultramarina.

No começo do século VIII a.C. os celtas foram para a península e se sobrepuseram a seus habitantes indígenas. Pouco a pouco, manifestou-se a supremacia da cultura céltica, que hegemonizou a cultura. Havia traços célticos comuns a quase todas as tribos.

No século III a.C., levados por circunstâncias políticas, os romanos envolveram com suas invencíveis legiões a Ibéria. A península estava dividida em dezenas de tribos das mais variadas origens, separadas por ódios enraizados e intransponíveis, oriundas do norte da África, emigradas do norte da Europa. Muitas delas estavam fortemente celtizadas, a ponto de adotarem muitos traços culturais dos celtas, inclusive a língua. A par disso, havia gente de procedências várias, que a guerra, o comércio e o espírito aventureiro lançaram sobre a península, como gregos e cartagineses, que muito antes dos romanos circulavam pela Ibéria. Os romanos procuraram unificar as tribos, mas até mesmo por causa de condições geográficas e étnicas, a Hispânia ficou sempre dividida em três províncias. Outras conquistas incorporaram novos povos à península. Os romanos modificavam o menos possível as unidades territoriais que encontravam, e respeitavam subdivisões étnicas e geográficas.

A implantação do latim na Península Ibérica constituiu fator decisivo para a formação da língua portuguesa. O prestígio dos romanos – invencibilidade das armas e traços da civilização helênica – impunha-se aos povos conquistados. Houve três fases da conquista romana na Ibéria: a expectativa dos povos conquistados (culturas frente a frente), a marginalidade (homens que participam das duas culturas – bilíngües) e vitória da cultura romana por todo o território, sobretudo impulsionada pelo cristianismo. A rápida difusão do latim relaciona-se ao prestígio dos conquistadores. O latim era a língua oficial, então, nem um documento era escrito ou transmitido em

língua indígena. As inscrições das moedas de circulação eram em latim. As novas técnicas tinham vocabulário latino. Soldados e comerciantes foram grandes propagadores do latim. Com as coisas dos negociantes iam as palavras que as nomeavam. O teatro foi outro importante veículo de propagação. Com exceção dos bascos, os povos da península adotaram o latim como língua e se cristianizaram. Na parte da Lusitânia, a romanização penetrou lentamente.

O traço mais saliente do estrangeiro é a língua, por isso, nas cidades, para se obter a cobiçada cidadania romana, precisava-se dominar o latim. Nas escolas, os indígenas adquiriam a mentalidade de um romano. Os romanos foram fundadores e difusores de escolas. O gramático, como conservador do idioma, encarregado de preservar a língua, tornou-se o mais eficaz representante do espírito romano.

Os romanos construíram estradas, que levaram a cultura romana a tribos longínquas, pois facilitavam a comunicação das províncias entre si e com o império. Sob a nova língua e novos costumes viviam antigas dissidências e tradições. Com o passar dos séculos, os estratos sociais e lingüísticos, por via da criação dos novos reinos, tiveram de ser reintegrados e se estabeleceu uma nova norma, que aceitava os fatos consumados da complexa estrutura lingüística da época.

O latim, obviamente, modificava-se. O acento, que era musical, passou a intensivo, pouco a pouco. Esse fato é de capital importância, pois o acento de intensidade conduz ao abreviamento e até à queda das vogais átonas, enquanto alonga a sílaba sobre a qual recai. Então, acarreta a subversão do sistema da quantidade silábica. No século III d.C., com os terríveis abalos que sofreu o império romano, o acento de intensidade triunfou. No latim, a última sílaba nunca era acentuada nos dissílabos e, nos polissílabos, o acento caía na penúltima sílaba. A fala corrente acentuava sempre a penúltima sílaba.

No início do século III, bárbaros germânicos invadiram a península. Como conseqüência, enriqueceu-se o vocabulário da região e iniciou-se a diferenciação lingüística entre partes da península.

No século VIII, povos árabes a invadiram e dominaram. Eram chamados de mouros pelos habitantes da península. O árabe era sua língua e sua religião era o

islamismo. Tanto a língua como a religião eram diferentes das praticadas na região, e não houve imposição de uma ou outra. A língua árabe era a oficial, mas o latim, já bastante diferenciado, era a língua de uso. Com os árabes vieram os judeus sefarditas, que também exerceram influência cultural na região. Não houve graves imposições aos godo-romanos, apenas a convivência com populações árabes. Ao longo do contato entre muçulmanos e cristãos, estes aceitaram numerosas palavras que designavam objetos de uma nova moda, aquisição ou descoberta, como as relacionadas ao militarismo (“almirante”, “arsenal”), a casa (“alicerce”, “alcova”), a ferramentas (“alicate”), ao campo (“limão”, “alfazema”, “alqueire”, “arroz”), dentre tantas outras (“xarope”, “álgebra”). São todos nomes concretos e em muitos deles está o artigo árabe invariável *al*.

Dos séculos passados entre a chegada dos bárbaros e a dos árabes à península não ficaram documentos lingüísticos, no entanto, é certo que o latim se transformou nesse tempo.

Desde os tempos romanos, a faixa ocidental da Hispânia apresentava características que a extremavam do resto do território, sobretudo a região centro-norte, onde se formaria o condado de Castela, cujo destino conduziria a hegemonia de quase toda a península.

No reinado dos reis católicos da Espanha, Fernando e Isabel, encerrou-se o período de dominação dos árabes, que durou sete séculos. As lutas entre cristãos e mouros pode ser dividida em três fases: a defensiva (720 a 1002, por iniciativa dos mouros), transição (até 1045, em que os cristãos começaram a reagir) e a da reconquista (até 1250), que marca o recuo dos muçulmanos.

A nação portuguesa constituiu-se no ano de 1128, na Batalha de São Mamede, sob a liderança de Afonso Henriques. Foi a vitória de uma nova geração. Em 1139-40, Dom Afonso Henriques selou a independência da faixa ocidental da península – *Portucale* –, que pertencia ao reino de Castela. Iniciou-se o Estado de Portugal como nação.

Com a evolução dos dialetos românicos, definiram-se três grupos lingüísticos, no século XII: o galego-português, o catalão e o castelhano. Quando Portugal separou-

se da Galiza, falava-se galego-português em toda a região da Galiza e da jovem nação portuguesa. O português originou-se, assim, do galego-português medieval.

No início do século XIII, surgiram os primeiros textos redigidos em galego-português, empregado em toda a Península Ibérica como veículo das cantigas trovadorescas que ali floresceram e também em forma de prosa, em documentos. Com a independência de Portugal, fatores políticos, econômicos e sociais determinaram a quebra da relativa unidade lingüística galego-portuguesa. Já separado do galego por uma fronteira política, o português, bastante diferenciado dos outros falares da região, seguiu seu curso, tornando-se a língua de Portugal. Então teve início a fase histórica do português, com a constituição da nova nacionalidade.

É na base da língua escrita que se costuma considerar, para a língua portuguesa, o período arcaico (até o século XV) e o período moderno.[2] A partir do século XV, a expansão das navegações encarregou-se de levá-la a outros horizontes. Assim, já com escrita, o português veio para o Brasil. Os negros escravos que eram trazidos para cá, em sua maioria, vinham da África e de Cabo Verde, e falavam português, embora bastante simplificado e, por isso, diferenciado.

A história da ortografia da língua portuguesa pode dividir-se em três períodos: fonético, até o século XVI; pseudo-etimológico, desde o século XVI até 1911; e moderno, desde 1911 até hoje, que se refere à ortografia simplificada.

Quando a língua portuguesa começou a ser escrita, quem a escrevia procurava representar foneticamente os sons da fala.[3] Essa representação, no entanto, não era satisfatória. Por um lado, não havia norma e, assim, por exemplo, o som /i/ podia ser representado por "i", por "y" e até por "h", a nasalidade, por "m", por "n" ou por til. Por outro lado, a ortografia conservou-se em certos casos antiquada em relação à modificação da pronúncia das palavras, como em *leer* ("ler") e *teer* ("ter"). Nos

[2] Foi somente no fim do período arcaico que os vocábulos proparoxítonos entraram na língua portuguesa, por empréstimo do latim clássico pelo modelo italiano.

[3] Numa ortografia fonética, a cada som corresponderia uma letra ou grupo de letras únicos e a cada letra ou grupo de letras, um som único. Ainda, pelo menos no caso das línguas indo-européias, seria assinalada de algum modo a sílaba tônica (CARVALHO, 1996).

documentos mais antigos, de qualquer modo, observa-se a procura por uma grafia fonética.

Eis um exemplo de prosa coloquial em português medieval, do século XIV, de autor anônimo:

> Grandes graças devemos dar a nostro senhor que non quer que peccadores se perçã, mais que sse tornẽ a ele e que façã peendença. Pois ouvide hun miragre que vos eu contarei que aves em nosso tẽpo. A mĩ, Jacobo, semelhou assi que escrevesse aos sanctos homẽes algũas cousas boas e que leendo-as e ouvyndo-as as creara e quefilhẽ ende conforto pêra sas almas.
> O misericordioso Deos parou assi este mũdo que o que fezer corregimentõ de seos peccados que aja ende perdon, ca no outro nõ averia seno direito jayzo ca receberá cada huu segundo sas obras. Pois ouvyde-me agora com todo amor e com toda entença de vosso coraçõ, ca esto que eu vos quero contar de door he de vossos pecados (*apud* SILVA NETO, 1988, p.400).

Da mesma época, este é um exemplo de prosa considerada refinada, de Fernão Lopes:

> Emtom começarom damdar, e passada a pomte chegamdo aa coyraça, chamou o Iffamte huum dos seus, e disse: "Vos sabees esta cidade, e as entradas e sahidas della, melhor que outro que aqui vaa, por que estevestes ja aqui no estudo: Dona Maria pousa nas casas d'Alvaro Fernandez de Carvalho, emcaminhaae per tal logar, per hu possamos hir a ellas, mais apressa e fora de praça que seer poder". E el respomdeo que asi o faria: e emtom os levou aa Igreja de Sam Bertolameu, domde naçe huuma estreita rua, que dereitamente vay sahir aas portas daquellas casas [...] (*apud* SILVA NETO, 1988, p.400).

Com o decorrer do tempo, essa simplicidade desapareceu por causa da influência do latim. Apareceram grafias como *fecto* ("feito"), *regno* ("reino"), *fructo* ("fruto"). Uma das características da Renascença (século XV e XVI) foi a admiração pelos tempos clássicos e, em particular, pelo latim. Isso consolidou e levou ao extremo a influência daquela língua na escrita do português. Disso resultou o aparecimento de muitas consoantes duplas e dos grupos *ph*, *ch*, *th* e *rh*, cujo uso antes era raro. Por outro lado, já nesse tempo, havia disparates, tais como *lythographia*, *typoia*, *lyrio*. Por essa razão, chama-se pseudo-etimológico ao período em que essa tendência se impôs.

Vieram reações no sentido de simplificar a ortografia, mas, apesar disso, na quase totalidade dos escritos, procurava-se a grafia mais complicada. Naquela época, o número de acentos era restrito e empregado com finalidades diferentes das atuais. No decorrer do século XIX, começou-se a compreender a falta de justificativa de muitas das grafias complicadas que então se usavam, contudo, simplificou-se em demasia. Como resultado, no fim do século XIX, a desordem ortográfica era total. Cada um escrevia como lhe parecia melhor.

No Brasil, a ortografia "oficial" era como no texto tipográfico a seguir, que é uma resolução oficial do Imperador Dom Pedro II enviada ao comando militar da Província de Santa Catharina, em 1824, segundo ano da Independência:

> 1824 25 de Maio
>
> DOM PEDRO PELA GRAÇA DE DEOS, E / Unanime Acclamação dos Povos, Imperador Constitucio- / nal, e Defensor Perpetuo do Imperio do Brasil. Fa- / ço saber aos que esta Minha Provisão virem; que / Tomando em Consideração, O que Me foi presente em / Consulta do Conselho Supremo Militar de seis de Ou- / tubro do anno passado, acerca do abuso com que as Juntas dos Governos Provizorios de differentes Provin- / cias deste Imperio, com transtorno do Serviço, tem con- / ferido Patentes de Commissão a muitos Officiaes; e Con- / formando-Me inteiramente com o Parecer do Conselho: / Hei por bem por minha Immediata, e Imperial Re- / solução de vinte e seis de Abril proximo passado, De- / clarar, que taes Patentes só poderão ter lugar, por Im- / mediato Despacho Meu. As Authoridades, e pessoas / quem o cumprimento desta pertencer o cumprão, e / guardem tão inteiramente, como nella se contem. O / Imperador o Mandou, pelos Conselheiros de Guerra / abaixo assinados, ambos do seu Conselho. José Rebel- / lo de Souza Pereira a fez no Rio de Janeiro aos vin- / te e cinco dias do mez de Maio anno do Nascimento de Nosso Senhor Jesus Christo de mil e oitocentos / vinte e quatro. O Conselheiro João Valentim de Faria Souza Labatto, Secretario de Guerra a fiz escre- / ver, e sobscrevi. /
>
> *Rodrigo Pinto Guedes Joaquim d'Oliveira Alvares./*
>
> Por Immediata, e Imperial Resolução de S. M. o / Imperador de 26 de Abril de 1824. /
>
> Registada a fl. 48 vers. Do Livro 1.º de Provisões. / Secretaria do Conselho Supremo Militar 31 de Maio / de 1824./
>
> *José Maria da Cunha Cabral/*
>
> Na Imprensa Nacional. //
>
> (GRÉGGIO, MERKLE e VASILÉVSKI, 1999, p.99)

Esse texto pode ser descodificado facilmente por um falante nativo alfabetizado do português contemporâneo. Ele é menos fonético do que seria hoje (*Deos*, *taes*, *officiaes*), por um lado, e mais fonético, por outro (*provizorios*). Chama a atenção a pouca acentuação gráfica (*unanime*, *proximo*, *Imperio*, *perpetuo*, *provincias*, *secretario*), a geminação (*differente*, *nella*, *anno*, *acclamação*, *commissão*, *officiaes*, *immediata*) e as grafias *cumprão*, *mez*, *authoridades* e *Christo*. Isso mostra que as mudanças ao longo do tempo são lentas e poucas.

Em 1911, o governo português nomeou uma comissão para estabelecer a ortografia a usar nas publicações oficiais. Dessa comissão fazia parte o foneticista Gonçalves Viana (1840-1914), que em 1904 apresentara um projeto de ortografia simplificada. A comissão praticamente adotou o que propunha Gonçalves Viana, e a nova ortografia foi oficializada por portaria em 1º. de setembro de 1911. Essa reforma da ortografia, a primeira oficial em Portugal, foi profunda e modificou completamente o aspecto da língua escrita, aproximou-o muito do atual. Foi uma mudança verdadeiramente radical e feita sem acordo com o Brasil. No Brasil, importantes filólogos adotaram-na, como José Oiticica, Antenor Nascentes, Mário Barreto e Souza da Silveira, mas outros não.

Ao fazer desaparecer muitas consoantes dobradas e os grupos *ph*, *th*, *rh* etc., a reforma eliminou exageros do período pseudo-etimológico. Essencial da reforma ortográfica de 1911 foi acabar com o monopólio da etimologia, ao aproximar a ortografia oficial de uma escrita fonológica. Apesar de tudo, fizeram-se vastas concessões a hábitos anteriores, com um ou outro pretexto, como manter diversas consoantes mudas (“homem”, “directo”, “sciência”). Um ponto em que a reforma se afastou da tradição dos primeiros tempos do português escrito foi a introdução profunda de acentos. Em particular, passaram a ser acentuadas todas as palavras esdrúxulas, o que não acontecia antes.

A seguir à reforma de 1911, houve vários ajustes efetuados por portarias de 1920, 1929 e 1931. A grande reforma seguinte foi resultante do acordo ortográfico entre Portugal e Brasil de 1945, a qual, levemente alterada por um decreto de 1971,

deu origem à ortografia oficial que até agora se usa em Portugal. O acordo de 1945 anulou algumas modificações introduzidas em 1911 e 1931.

No Brasil, no século XIX, a ortografia estava no mesmo estado que em Portugal. Havia unidade no caos. Em 1907, a Academia Brasileira de Letras estudou um projeto de reforma análogo ao de Gonçalves Viana, que levou à reforma portuguesa de 1911. Esse projeto baseava-se no do foneticista português, mas nele colaboraram vários brasileiros ilustres, como Euclides da Cunha e Rui Barbosa.

Tal projeto não foi adiante e Portugal avançou sozinho para a reforma. Assim, e apesar de a reforma portuguesa ser defendida sem alterações para uso no Brasil por filólogos brasileiros do calibre de Antenor Nascentes e Mário Barreto, durante alguns anos, ficaram os dois países com ortografias completamente diferentes: Portugal com uma ortografia moderna, o Brasil com a velha ortografia pseudo-etimológica.

Em 1924, as duas academias, a Brasileira de Letras e a das Ciências de Portugal, resolveram definir uma ortografia comum. Para isso, o Brasil teria de se aproximar de Portugal. Houve em 1931 um acordo preliminar entre as duas academias, em que se adotava praticamente a ortografia portuguesa. Tal acordo deixou a ortografia similar à que se usa hoje no Brasil, mas Portugal não o seguiu na prática. Esse acordo dizia que não se escrevem consoantes que não se pronunciam, mas em Portugal continuou-se a escrever “acto”, “adoptar”.[4]

Na década de 1940, os vocabulários que se publicaram, o da Academia das Ciências de Portugal (1940) e da Academia Brasileira de Letras (1943), continham divergências. Por isso, houve, ainda em 1943, em Lisboa, uma convenção que deu origem ao Acordo Ortográfico de 1945. Esse acordo tornou-se lei em Portugal, mas no Brasil não foi ratificado pelo Congresso Nacional. Por isso, os brasileiros continuaram a regular-se pela ortografia do Vocabulário de 1943.

Em 1971, novo acordo oficializado entre os dois países aproximou um pouco mais a ortografia do Brasil da de Portugal. Ambas as grafias eram, a essa altura, razoáveis, com algumas diferenças. Em 1973, recomeçaram as negociações e, em

[4] Gonçalves Viana criou a grafia dessa consoante não pronunciada para marcar o timbre aberto da vogal anterior (CRUZ, 2007).

1975, as duas academias mais uma vez chegaram a um acordo, o qual não foi transformado em lei, devido a circunstâncias adversas que não permitiram consideração pública da matéria.

Em 1986, o presidente brasileiro José Sarney tentou resolver o assunto, que havia longo tempo se arrastava, e promoveu o encontro dos sete países de língua portuguesa no Rio de Janeiro. Desse encontro, mais uma vez, saiu um acordo ortográfico que, mais uma vez, não foi adiante, devido a surpreendente alarido que se levantou em Portugal, sobretudo por parte dos adversários da união ortográfica.

Em 1990, os responsáveis portugueses pelo acordo de 1986, que garantia unificação quase total da ortografia da língua, produziram um acordo que não a unificava e tinha grafias duplas desnecessárias.

Em 2007, o acordo entre os sete países foi foco de atenção. Pelo tratado, modificam-se em torno de 0,5% das palavras atualmente usadas no português do Brasil e 1,5% das utilizadas em Portugal. No Brasil, deixa-se de acentuar "idéia" e abole-se o trema. Os portugueses tiram o "c" não pronunciado de "acto" e o "p" não pronunciado em "óptimo". As regras para a utilização do hífen são unificadas, e ficam consagradas algumas diferenças: os portugueses tiram o "p" que não pronunciam de "recepção" e mantêm o "c" de "facto", porque "fato" em Portugal significa roupa. O alfabeto oficial, que tem 23 letras, passa a ter 26 letras, com a inclusão de *k*, *w* e *y*, que serão usadas em casos especiais.[5]

Atualmente, o português é uma das poucas línguas do mundo ocidental faladas por mais de cinqüenta milhões de pessoas e que têm mais de uma ortografia oficial. O castelhano, por exemplo, apresenta dezenas de variações de pronúncia na Espanha e América hispânica, mas apenas uma ortografia.[6]

O texto original do Acordo Ortográfico de 1990 previa a elaboração, até janeiro de 1993, de um vocabulário ortográfico comum, tão completo quanto desejável e tão normalizador quanto possível, no que se refere às terminologias científicas e técnicas.

---

[5] A pronúncia brasileira, em geral, repousa sobre um sistema fonético antigo e de aspecto urbano, por conta da distância de Portugal.

[6] A língua inglesa apresenta variações ortográficas nacionais comparáveis às observadas nas variedades nacionais da língua portuguesa, mas não há regulamentação oficial.

A Academia Brasileira de Letras publicou uma obra com 350 mil entradas em 1998,[7] o VOLP, mas a Academia das Ciências de Lisboa, com seu dicionário de 70 mil entradas, de 2001, está longe desse objetivo.

Críticos sustentam que a proposta tenta resolver um não-problema, uma vez que, apesar das diferenças ortográficas, as variantes escritas da língua portuguesa são perfeita e confortavelmente inteligíveis por seus leitores. O sucesso de vendas do escritor luso José Saramago no Brasil, cujos livros usam a grafia lusitana, é tomado como evidência disso.[8] Esses críticos apontam que as dificuldades de compreensão, quando ocorrem, são devidas às diferenças de vocabulário, que não dependem de ato normativo.

Portugal é o país que mais hesita na aceitação do acordo, porque esse tratado seria uma "abrasileiração" da escrita e a variante lusitana da língua ficaria prejudicada. É um sentimento esperado, uma vez que o nome da língua está no nome desse país, onde ela formou-se. No entanto, foram os portugueses que a trouxeram para o Brasil e a nacionalizaram por todo o imenso território brasileiro, à custa do extermínio de línguas indígenas, principalmente, e africanas.[9]

Mesmo assim, a nação portuguesa avançou na 12ª reunião do Conselho de Ministros da Comunidade dos Países de Língua Portuguesa (CPLP), em Lisboa, quando se dispôs a aprovar o protocolo modificativo do acordo ortográfico da língua portuguesa até o final de 2008. O conselho de ministros português posicionou-se favoravelmente ao acordo em março de 2008, quando aprovou uma proposta de resolução sobre o segundo protocolo modificativo ao acordo ortográfico da língua portuguesa, que foi ratificado em 2004 pela CPLP. Os demais países que ainda não o fizeram comprometeram-se a ratificá-lo logo.

---

[7] Talvez por conta desse acordo, o VOLP de 1998 da ABL está esgotado há anos, e não há nova versão nem reimpressão para compra, apenas está disponível para consulta na Rede Mundial.

[8] Isso se torna peculiar, já que nas academias brasileiras professores ensinam a futuros professores que Brasil e Portugal falam línguas diferentes. Ocorre que os brasileiros lêem os textos de Saramago 'sem sotaque'. Do ponto de vista escrito, esse fato mostra que as diferenças são irrelevantes para caracterizar outra língua. Do ponto de vista fonológico, talvez seja outra língua, talvez um dialeto complicado, na visão dos brasileiros.

[9] Sobre essa questão ver, por exemplo, Vasilévski (2004b).

Apesar disso, as editoras portuguesas ameaçam não adotar em seus livros as alterações previstas, por considerar que se trata de facilitar a entrada das editoras brasileiras nos países africanos, dentro dos quais as editoras portuguesas já estão, complementa-se. Apesar disso, prevê-se que o acordo entre em vigor em Portugal nos próximos seis anos. No Brasil, o acordo entrará em vigor em no máximo dois anos.

Diante dessas circunstâncias, cabe comparar geograficamente as nações que falam português no mundo:

Figura 1: Países onde se fala a língua portuguesa
Fonte: Adaptado de Wikipédia, 2007.

O mapa mostra até onde chegou a língua portuguesa e deixa clara a proporção em que se fala o português do Brasil – português com açúcar – em relação ao das outras nações. Mesmo ao se somar todos os outros países, o Brasil permanece como maior coletividade de língua portuguesa, e as outras juntas nem se aproximam de sua metade. O Brasil tem em torno de 184 milhões de falantes de português dos 210 milhões estimados que há no mundo.

Quando o homem conseguiu representar a língua graficamente e, então, pode deixar registros, começou a História, por volta de 4000 a.C. A língua é uma instituição cujas modificações se ligam indissoluvelmente à história da coletividade que a emprega. A história da língua é a história dos homens que a falam – sua cultura e

tradições. A língua portuguesa formou-se ao longo do tempo a partir da mistura de línguas – principalmente latim e árabe – e interferência de muitas culturas: céltica, romana, germânica, muçulmana, dentre outras. Assim chegou ao Brasil, onde se enriqueceu com indianismos, africanismos e, mais recentemente, italianismos, germanismos e niponismos, dentre outras contribuições étnicas. Assim, "alface" é contribuição árabe, "abacaxi" é contribuição indígena, "clube" é contribuição inglesa, todas perfeitamente harmonizadas no sistema alfabético do português.

Esta pesquisa lida com essa riqueza e complexidade de tema fundamental, e, nesse sentido, enfoca as relações entre fala e escrita.

1.2 SISTEMA VERBAL

O sistema verbal de uma língua é o conjunto de suas modalidades orais e escritas.

A história esclarece que a linguagem verbal oral se desenvolve espontaneamente, desde que haja traços de humanização, enquanto a linguagem verbal escrita é uma invenção, cuja aprendizagem intensiva e sistemática se faz necessidade, na maioria dos casos (SCLIAR-CABRAL, 2003).

As línguas estão em perpétua mudança, embora somente o repouso seja facilmente perceptível. A evolução explica-se, principalmente, pela descontinuidade da transmissão e pela própria constância do uso. O movimento geral da evolução, observado historicamente, proporciona o estabelecimento de uma série de correspondências (ou leis) fonéticas, embora elas representem apenas um meio prático para investigações. Elas são tábuas guias através de espessas florestas. As diferenças de fase para fase são atenuadas pela língua escrita, que é conservadora e se submete às tradições literárias. Ainda assim, se ressalvados os períodos de decadência, nos quais se imita uma fase anterior, a própria língua escrita oferece diferenças de fase para fase (SILVA NETO, 1988).

A língua e a vida social modificam-se juntas. O que a criança realiza em quatro ou cinco anos, com base em modelos existentes, os ancestrais humanos levaram de

dois a três milhões de anos para realizar, no entanto, a ordem hierárquica segundo a qual se cumpre essa mudança deve ser ainda a mesma (MALMBERG, 1993). Ao cabo de seu aprendizado, a criança fixa uma língua que não é exatamente a mesma das pessoas que lhe serviram de modelo. Essa diferença, imperceptível numa geração, acumula-se aos poucos (SILVA NETO, 1988).

As palavras mais freqüentemente usadas são as que mais transformações sofrem. Grupos de palavras aglutinam-se e o desgaste provoca reações. Por isso, a todo instante surgem inovações, cujo destino depende da estrutura social, no caso, da força com que a língua, como instituição, impõe-se aos indivíduos, da aceitação coletiva (SILVA NETO, 1988).

Diferentemente de meios de comunicação como a pintura e o desenho, a escrita é recente, em comparação com a modalidade oral. O sistema sumério, considerado o mais antigo, apareceu há 5000 anos. Foi necessário muito tempo para acumular conhecimentos e tecnologia suficientes e pressão de necessidades socioeconômicas, para se descobrir que as palavras escritas se constituíam por unidades menores do que a sílaba, responsáveis pelas diferenças de significados, e que essas pequenas unidades poderiam ser representadas por signos gráficos (SCLIAR-CABRAL, 2003).

Os agrupamentos dos fonemas em série (correlações) são as expressões de um esforço de economia que garante efeito máximo, por meio de esforço mínimo (esforço medido pelo mínimo de distinções a dominar). Segundo a teoria evolutiva[10] de Martinet, toda língua procura garantir a economia de seu sistema e as mudanças realizadas são condicionadas por necessidade de economia, contudo, muitas vezes, ocorrem lacunas que os falantes procuram preencher, inconscientemente (MARTINET, 1955 *apud* MALMBERG, 1993).

A língua falada, que se faz acompanhar por gestos, por entonação de voz e pela variada expressão facial, é rica em exclamações, pleonasmos e redundâncias de toda ordem. A língua escrita não dispõe desses elementos, mas se dirige a ausentes cujas

[10] A evolução lingüística não é apenas um fato de mudança fonética e fonológica, no entanto, as modificações freqüentemente começam como modificações de pronúncia. As distinções enfraquecem e desaparecem. Palavras e formas coincidem, introduzem-se novas palavras para evitar homônimos e manter a independência dos signos (MALMBERG, 1993).

reações não observa, portanto, é mais comedida. A língua falada reflete a expressão de pessoas cujo saber é tradicional e adquirido na escola da vida. A língua falada é anônima. O problema vital da língua escrita são suas relações com a língua falada (SILVA NETO, 1988).

Considera-se língua escrita

> O sistema de meios gráficos empregados com o propósito de produzir enunciados [e textos] escritos aceitáveis numa dada comunidade lingüística. Tais meios incluem não apenas os grafemas[11] (implementados pelas letras), mas também as marcas diacríticas,[12] compartilhando com os grafemas seus lugares segmentais[13] nos enunciados escritos, [e] os meios estabelecidos de combinar mutuamente tais grafemas (as leis que governam esta combinação dos grafemas são muitas vezes designadas como regras grafotáticas)[14] (VACHEK, 1973 *apud* SCLIAR-CABRAL, 2003, p.27).

Uma ou mais letras (grafemas) representam os fonemas e alguns de seus alofones, que resultam nas unidades que distinguem o significado na escrita (a segunda articulação). Outro princípio divergente também ocorre: o etimológico (SCLIAR-CABRAL, 2003).

As pessoas pertencentes a uma mesma comunidade intercambiam mensagens orais inteligíveis, independentemente de variedades determinadas pelo contexto e por vários outros fatores. Da mesma forma ocorre na língua escrita, em que um texto é decifrado por uma comunidade graças aos princípios de reconhecimento da palavra, pois os valores atribuídos pelos membros dessa comunidade aos signos são os mesmos (SCLIAR-CABRAL, 2003). Nesse caso, a variação é menor, pois a escrita é fixa e

---

[11] Grafema é uma ou duas letras (dígrafo) que representam um fonema (SCLIAR-CABRAL, 2003). Excepcionalmente, três fonemas podem ser representados por uma letra, como o "x" de "tórax", cuja transcrição pode ser /tʰɔrakiS/.

[12] Marcas diacríticas são sinais apostos às letras para modificar-lhes o valor, como o cedilha, o til, o trema, os acentos agudo e circunflexo (SCLIAR-CABRAL, 2003).

[13] Lugares segmentais são lugares que correspondem aos fonemas, como o "sc", que corresponde ao lugar segmental do fonema /s/ em "nasce" (*Idem*).

[14] Regras grafotáticas determinam as combinações possíveis dos grafemas em um dado sistema alfabético, os valores decorrentes dos contextos que os cercam e as posições possíveis e vedadas que podem ocupar. Por exemplo, numa mesma sílaba, pode haver a combinação de "t" e "r" apenas nessa ordem (*Ibidem*).

permanente em relação à fala. A escrita é um registro, é uma história contada de um único modo, apesar de ser fiel a sua época.

Os diferentes sistemas escritos revelam a concepção que seus inventores tinham sobre como estão estruturadas as línguas (SCLIAR-CABRAL, 2003).

As gramáticas escolares focalizam explicitamente a escrita. O estudante vai para a escola falando satisfatoriamente a linguagem familiar. A técnica da língua escrita ele tem de aprender na escola. Ocorre que a língua escrita manifesta-se em condições muito diversas das circunstâncias da língua oral. Por isso, tantos estudantes psiquicamente normais, que falam bem no intercâmbio de todos os dias, são desoladores na escrita. A fala se desdobra numa situação concreta, e isso desaparece na escrita. Essa é uma profunda diferença entre os dois sistemas verbais (CÂMARA JR., 1986).

A escrita não reproduz fielmente a fala. Ambas têm leis e caminhos próprios. Do ponto de vista sociológico, a língua escrita sobrepõe-se à língua oral, pois rege toda a vida geral e superior do país (CÂMARA JR., 1986). Ainda, somente depois de dominar a fala pode-se aprender a ler e escrever (JAKOBSON, 1969 *apud* CÂMARA JR., 1986). Isso impõe a tarefa de fazer a descrição lingüística em função da língua oral.

Nenhum indivíduo sabe 100% de sua língua. Tanto sua força quanto sua fraqueza fazem parte de suas possibilidades lingüísticas e devem, conseqüentemente, fazer parte de uma transcrição geral do comportamento lingüístico humano. A língua é, ao mesmo tempo, o que o indivíduo consegue e o que não consegue fazer: êxito e fracasso (MALMBERG, 1993).

1.3 FONÉTICA E FONOLOGIA

As relações entre as modalidades escritas e orais de uma língua são objeto de estudo da fonética e da fonologia. Enquanto a fonética preocupa-se com descrever todas as línguas de forma detalhada – pois estuda os sons da fala, os fones –, a fonologia estuda os fonemas de uma língua. Assim, a fonologia provê uma transcrição

geral de uma língua, que engloba o máximo de variações possível. A fonologia estuda a invariabilidade profunda, enquanto a fonética abrange variações superficiais. A fonética envolve muito mais símbolos do que a fonologia, afinal, para cada variação de um fonema, há um símbolo distinto em fonética, mas em fonologia é o mesmo símbolo, pois ainda se trata do mesmo fonema, uma vez que a variação não muda o significado da palavra em que ele é usado.

Fonologia é a análise funcional do emprego que uma língua faz de seus recursos sonoros. A fonologia faz abstração das qualidades físicas dos elementos, reservadas ao domínio da fonética pura. Em sentido estrito, a fonética é o estudo da substância da expressão (realização dos sons em sua multiplicidade e variação) e a fonologia, o da forma (relações, classes, natureza abstrata, que se realiza na substância) (MALMBERG, 1993). Apesar disso, na prática, diferentes autores têm diferentes concepções acerca dos limites das duas disciplinas, e isso se revela nas transcrições que apresentam em suas publicações. Então, como ocorre nesta pesquisa, deve-se assumir uma postura própria e coerente em relação a isso.

A partir disso, as transcrições [dʃi] e [tʃi] são fonéticas, pois correspondem a variantes de /di/ e /ti/, que são transcrições fonológicas, gerais, portanto. As duas disciplinas estão internamente ligadas, tanto que, algumas vezes, há controvérsia entre os teóricos quanto a certa transcrição ser fonética ou fonológica.

Cabe lembrar que, quanto à simbologia, transcrições fonéticas vêm entre colchetes, como [tʃˈia], e transcrições fonológicas vêm entre barras, como /tˈia/.[15]

1.3.1 Fonema

Às classes dos sons da língua dá-se o nome de fonema. Qualquer idioma os tem em número considerável, e as letras do alfabeto não bastam para representar os fonemas na escrita. Para remediar a deficiência, recorre-se a certas combinações, como

[15] Todas as transcrições deste estudo foram feitas no programa Nhenhém (2008). Essa foi a primeira prática do aplicativo. O Nhenhém posiciona a marca de tonicidade dos vocábulos antes da vogal tônica, por motivo que será oportunamente esclarecido.

munir algumas letras de sinais complementares (acentos gráficos nas vogais) e juntar duas letras para denotar um fonema (“lh”, “nh”, “ch” etc.) (SAID ALI, 1964).

Foi a partir do estruturalismo que a língua passou a ser interpretada como um sistema, em que os fonemas são elementos da matéria fônica de um vocábulo e caracterizam-se pela circunstância de não se confundirem entre si (CÂMARA JR., 1977). Então, cada fonema tem pelo menos uma propriedade de seu feixe de propriedades que o distingue de todos os demais. Nesse sentido, traço pertinente ou distintivo de um fonema é toda característica fônica suscetível de diferenciar por si o significado de uma palavra ou frase. Traços pertinentes asseguram a função distintiva e devem ser as unidades básicas da fonologia.

Dessa forma, os sons não devem ser considerados fonemas, mas sim realizações de fonemas. Fonemas são abstrações psíquicas da língua. Trata-se de depreender os traços distintivos da fonêmica de uma língua, dentre todos os que a fonética faz conhecer, a partir de termos articulatórios (CÂMARA JR., 1977). Enquanto um fonema caracteriza um conjunto único de propriedades distintivas, o som é um conjunto de características distintivas e indistintivas, um símbolo material do fonema. Assim, cada som apresenta traços pertinentes de um fonema de que é realização e também uma série de outros traços fônicos irrelevantes, que dependem de diversas causas. Disso se tem que um mesmo fonema pode ser realizado por diversos sons, desde que estejam presentes os mesmos traços diferenciais.

Ao falar uma língua estrangeira, o maior problema é reproduzir os verdadeiros traços distintivos dos fonemas dessa língua (CÂMARA JR., 1986). A língua estrangeira pode ser mais rica ou mais pobre em unidades distintivas e em possibilidades sintagmáticas. Se a língua a se aprender tiver maior número de elementos fonológicos (fonemas, acentos) do que a língua materna, a dificuldade de aprendizado será maior. Em contrapartida, se a língua materna tiver mais desses elementos, há risco de o aprendiz introduzir na língua alvo traços que não pertencem a ela (MALMBERG, 1993).

A classificação dos traços distintivos pode ser feita por vários critérios. Ela pode ser baseada nas vibrações das ondas sonoras do ar quando se fala (acústica), pode

tomar como critério o efeito que o som emitido produz no ouvido humano (auditiva), também pode se dar pela definição dos traços pelo movimento dos órgãos fonadores (articulatória).

Os fonemas são a base elementar da língua, servem para distinguir as expressões dos signos e, conseqüentemente, para manter os signos separados, caracterizados e diferenciados, sendo essa precisamente sua função distintiva (MALMBERG, 1993).

A fonologia estuda os fonemas, que são os sons vocais elementares considerados a partir de suas propriedades articulatórias e acústicas relevantes – ou seja, seus traços distintivos –, e podem ser vogais e consoantes. Os fonemas são a divisão mínima da segunda articulação, entendidos como um conjunto de características distintivas, que os opõem entre si, de forma que cada fonema seja único e concorrente. Na realidade física, a emissão de sons é um contínuo (CÂMARA JR., 1986).

Fonema é um conceito da língua oral e não se confunde com a letra na língua escrita, na qual um mesmo fonema pode ser representado por diferentes letras (CÂMARA JR., 1986), como /ʃ/, que pode ser representado por “ch” ou “x”.

1.3.2 Alofone

Na segunda articulação, o fonema invariante se desdobra em alofones. Uma vez que o fonema é um feixe de traços distintivos e não distintivos (sons elementares), os segundos criam os alofones, pois um fonema abrange vários sons elementares, que podem ser percebidos como diferentes por um falante nativo da mesma língua. Por não haver troca de traço distintivo, os alofones podem causar estranheza a um ouvido não acostumado com tal pronúncia, mas não prejudicam o entendimento de o que é dito nem substituem uma forma da língua por outra. No entanto, podem causar variações ortográficas (CÂMARA JR., 1986). Os alofones, que correspondem às realizações de um fonema, não distinguem significado.

Quando as variantes de fonema dependem do ambiente fonético em que o som vocal se encontra, assimilam-se os traços dos outros sons contíguos. Eles afrouxam-se ou mudam-se as articulações, em virtude da posição fraca em que o fonema se acha. É o que ocorre na posição átona em português, sobretudo na posição final. Esse tipo de alofone chama-se posicional. Os alofones posicionais dão o sotaque local da fala (CÂMARA JR., 1986).

Alguns alofones são condicionados pelo contexto fonológico, como cada uma das vogais átonas portuguesas, em face da correspondente vogal tônica. Outros são variantes livres, determinadas pela flutuação que impera na língua em referência à realização fônica do fonema. Na fonologia, o que distingue o fonema do alofone é a capacidade do fonema de distinguir os significados da língua (CÂMARA JR., 1986).

1.3.3 Neutralização e Arquifonema

Ocorrem em português casos em que, em certas posições na sílaba, há neutralização das oposições distintivas entre duas vogais ou consoantes. Disso se origina o que se chama arquifonema. Então, arquifonema seria a representação do conjunto das propriedades distintivas comuns a dois fonemas que estão em oposição neutralizável em determinado ambiente.

As posições átonas favorecem esse fenômeno fonológico. O que interessa são as propriedades distintivas, pois diferenças até fisicamente muito grandes podem resultar na mesma coisa (CÂMARA JR., 1986). Sons vocálicos reduzidos são conseqüência da posição átona da vogal. Essencialmente, a redução do número de fonemas caracteriza as posições átonas, isto é, mais de uma oposição desaparece ou suprime-se, e fica para cada uma um fonema em vez de dois. É o que Trubetzkoy tornou um conceito clássico em fonologia com o nome de neutralização (CÂMARA JR., 1986).

Em português, neutralizam-se /e/ e /i/ e /o/ e /u/ em algumas posições átonas, sobretudo em final de vocábulo. Dessa forma, têm-se: “momento” → /momˈẽtu/ e “instante” → /ĩStˈãti/. Na pronúncia de sílabas travadas por “l”, que corresponderia ao fonema /l/, há neutralização desse fonema em favor de /w/, porque ele se comporta

como essa semivogal e, então, origina ditongo, na maioria dos falares brasileiros: “balde”, “beldade” → /bˈawdi/, /bewdˈadi/.

A letra “z” ora se descodifica como /s/ ora como /z/, de forma que desaparecem as oposições distintivas entre esses dois fonemas, o que origina o arquifonema |S|: “capaz” → /kapˈaS/, “capas” → /kˈapaS/. Vale observar que o som de /z/ é resgatado na frase, quando a ele sucede vogal: “capaz até disso” → /kapˈaz atˈɛ dˈisu/, “capas alaranjadas” → /kˈapaz alarãʒˈadaS/. Esse mesmo arquifonema ocorre em casos em que a consoante “s” pode copiar o traço de sonoridade do fonema consonantal seguinte a ela: “mastro” → /mˈaStru/, “esmago” → /eSmˈagu/.

Há também o arquifonema |R|, que representa a neutralização dos traços dos fonemas /R/ e /r/ em início de vocábulo, em final de sílaba e em início de sílaba interna não precedida de vogal (SCLIAR-CABRAL, 2003): “rigor”, “desregrar” → /RigˈoR/, /deSRegrˈaR/. Todavia, na frase, o fonema /r/ é resgatado, quando ocorre em sílaba final de vocábulo e precede vocábulo iniciado por vogal: “amor antigo” → /amˈor ãtˈigu/.

No arquifonema suspende-se a função do traço distintivo. Tanto no arquifonema |R| quanto no |S|, neutralizam-se os traços sonoro e surdo, por exemplo. Segundo Scliar-Cabral (2003), o arquifonema |R| cobre todas as realizações possíveis do grafema “r”, nas diversas variedades sociolingüísticas do português do Brasil. No entanto, para Carvalho (2008), o “r” inicial de “rato” → /ʁˈatu/ representa todas as pronúncias possíveis do “r” que inicia palavras, além de distinguir as pronúncias de “rato”, “pato”, “gato” etc., e isso é propriedade de fonema. Scliar-Cabral defende que, nesse caso, ocorre o fenômeno chamado lacuna de distribuição de fonema, por isso, não é possível par mínimo com /r/ e /ʁ/, independentemente de o que vier depois. Essa controvérsia é mote para prolongar essa discussão, a fim de amparar o uso do arquifonema |R|.

O uso de arquifonemas sempre foi problemático nos estudos da língua, desde sua introdução na fonologia por Trubetzkoy, pois há interpretações diferentes sobre o

que seria um arquifonema. Desse modo, sua aplicação deve ser comedida e coerente com alguns dos teóricos que dele tratam. Em alguns contextos, a noção de arquifonema não é produtiva, razão pela qual a fonêmica não gostava dessa interpretação.[16]

Originalmente, atribuía-se o valor de arquifonema somente a casos em que houvesse neutralização, ou seja, quando dois fonemas perdessem o valor opositivo em algum contexto. Assim, neutralização não é a simples não-ocorrência de um dos fonemas (como o caso de distribuição complementar,[17] que não cria um arquifonema, e o caso da vibrante como segundo elemento do aclive silábico: /brˈaza/, /frˈazi/ etc.) (CAGLIARI, 2002). Por esses motivos, é defensável considerar que, no início de palavras, não ocorre o arquifonema |R|, mas sim o fonema /ʁ/ (CAGLIARI, 2002; CARVALHO, 2008; CÂMARA JR., 1986).

> Em certas variedades do português (dialeto paulista e outros), ocorre uma oposição fonológica entre [x] e [ɾ], [ou /ʁ/ e /r/] quando [são] intervocálicos. Essa oposição [anula]-se em início de palavra, porque, nesse contexto, nunca se encontram ocorrências de [ɾ] [e não porque há perda de valor opositivo]. Veja os seguintes exemplos:
>
> | | | |
> |---|---|---|
> | | *carro* [kaxʊ] | *rato* [xatʊ] |
> | | *caro* [kaɾʊ] | *Rita* [xita] |
> | | *murro* [muxʊ] | *roda* [xɔda] |
> | | *muro* [muɾʊ] | *rumo* [xũmʊ] |
> | contexto: | V__V | #__V |
> | *status*: | oposição entre [x] e [ɾ] | [...] não-ocorrência de [ɾ] |
>
> (CAGLIARI, 2002, p.48)

Apesar disso, em outros contextos, a noção de arquifonema é útil, como em final de sílaba, em que podem ocorrer os alofones, porém, neutralizados em suas oposições. Nessa circunstância, "s" pode ocorrer como /s/ ou como /z/ – quando, então, é representado pelo arquifonema |S| –, mas não em oposição (CAGLIARI, 2002).

---

[16] Toma-se por base o conteúdo de correspondência eletrônica do professor Luiz Carlos Cagliari, enviada à pesquisadora em 22 de julho de 2008, em resposta à mensagem intitulada *Arquifonema eletrônico*.

[17] Distribuição complementar acontece quando um único fonema tem duas variantes, e cada qual ocorre em contextos diferentes. Ocorrência complementar acontece quando dois fonemas neutralizam-se, porque ocorrem em contextos diferentes (CAGLIARI, 2002).

Em comum, Scliar-Cabral (2003) e Cagliari (2002), apesar de adotarem diferente simbologia em seus trabalhos, bem como outros autores que tratam do tema, aceitam os arquifonemas |R| e |S|, no declive da sílaba, e isso respalda a adoção desses arquifonemas, unicamente nesse caso. Então, o "r" que inicia palavras, como em "real", é fonema, pois sua aceitação como arquifonema diverge entre esses autores e Câmara Jr. (1986) sustenta que, nessa posição, somente ocorre o fonema forte /ʀ/.[18]

1.4 SISTEMA ORTOGRÁFICO OFICIAL DO PORTUGUÊS DO BRASIL

O sistema ortográfico brasileiro oficial é composto pela representação das consoantes e vogais em letras (que se articulam em sílabas, que se articulam em palavras, que se unem em frases, que se unem em textos), diacríticos e sinais de pontuação. Grosso modo, vogais são sons que se pronunciam sem auxílio de outros sons e, pela etimologia, consoantes (com = junto, soante = que soa) são sons que somente podem soar com auxílio de uma vogal. No entanto, a fonética demonstra que, das consoantes, somente as oclusivas, também chamadas de momentâneas, plosivas ou obstruintes, e as constritivas vibrantes não podem ser articuladas sem apoio vocálico.

1.4.1 Vogal Oral

Vogal é um som produzido pela ressonância bucal onde a corrente de ar passa livremente. A realidade da língua oral é muito mais complexa do que dá a entender o uso aparentemente simples das sete vogais que se ensinam na escola, pois elas multiplicam-se em muitas variantes (CÂMARA JR., 1986). A vogal /a/ é considerada vogal fundamental, por ser a que primeiro se ouve quando vibram as cordas vocais e quando não se contrai nenhuma das partes da boca. É a que demanda menos esforço, e a criança a emitiria por primeiro (SAID ALI, 1964). Jakobson (1971, p. 21-30)

[18] Por causa de incompatibilidade entre fontes do IPA e o computador, como será posteriormente esclarecido, usa-se no Nhenhém e ao longo deste estudo o símbolo /ʀ̥/ para representar o fonema forte /ʀ/.

defendeu a idéia de que assim acontece em virtude do contraste máximo na sílaba /pa/, entre a consoante mais fechada e surda e a vogal mais aberta.

As vogais orais do sistema ortográfico oficial do português são representadas por cinco letras, com ou sem diacríticos: “a”, “e”, “i”, “o” e “u”. Todas as letras que representam as vogais podem receber o acento agudo, mas somente as letras “a”, “e” e “o” podem também receber o acento circunflexo e apenas “a” e “o” podem receber o til para marcar nasalidade. Toda a vez que a letra “o” receber o acento agudo, trata-se da vogal posterior baixa acentuada /ɔ/, porém, isso não é verdade em relação a “é”, porque, quando se tratar de oxítonos terminados em “em”, o acento agudo apenas marca a intensidade, e não a vogal anterior baixa acentuada /ɛ/. Pode-se também afirmar que o acento circunflexo no “e” dos verbos “ter” e “vir” e seus derivados, na terceira pessoa do plural do presente do indicativo, é uma marca morfossintática coesiva, e não assinala a diferença de timbre entre a vogal fechada e a aberta: coincide, redundantemente, com a marca de nasalização do “m” final. As duas letras “e” e “o” também podem ser descodificadas com timbre aberto, em ocasiões em que não recebem acento agudo, mas podem também representar as vogais /i/ e /u/, quando estiverem em posição átona, e as semivogais nasalizadas, como em “mãe” e “mãos”.

A língua espanhola, por exemplo, tem menos timbres vocálicos do que o português, que é muito mais variado nesse aspecto. Por isso, falantes de espanhol têm dificuldade de compreender o português falado, enquanto os falantes de português acompanham razoavelmente bem o espanhol. Comparado com o sistema vocálico do português, o sistema espanhol é simples (CÂMARA JR., 1997).

As vogais são centro de sílaba e podem ser tônicas e átonas, orais e nasais (ou nasalizadas). O acento constitui circunstância ótima para caracterizar as vogais portuguesas (CÂMARA JR., 1997).

Na escrita, em posição tônica, há sete vogais em português, as que se aprendem na escola. Em posição átona, a neutralização entre as vogais é favorecida, como ocorre com “e” e “o” em final de vocábulo não oxítono, que se lêem como /i/ e /u/, na maioria das variedades sociolingüísticas. Assim, nesse caso, elas apresentam timbre diferente do que apresentam quando estão em posição tônica. As distinções então tendem a

reduzir-se. A oposição entre /e/ e /i/ e entre /o/ e /u/ pretônicos é funcionalmente pobre, como nas pronúncias [kolˈɛʒiu] e [kulˈɛʒiu], mas há flutuação (CÂMARA JR., 1997).

Há, então, três quadros átonos distintos para as vogais portuguesas no Brasil: um para as vogais pretônicas, um para as vogais postônicas não finais (nos vocábulos proparoxítonos) e um quadro para as vogais átonas finais. Muitas vezes, há ilusória diferença gráfica. É nas posições átonas que se afirma a diferença de vocalismo entre Portugal e Brasil (CÂMARA JR., 1997). Em posição átona, as vogais "a", "e", "i", "o" e "u" da escrita podem ser reduzidas na pronúncia.

1.4.2 Vogal Nasal e Nasalizada

A língua portuguesa caracteriza-se, dentre as línguas românicas, por emissão nasal das vogais, muitas vezes. Na emissão sonora das vogais, mediante abaixamento do véu palatino, obtêm-se as vozes nasais. Vogal nasal seria, portanto, aquela em cuja emissão parte do ar é desviada para as fossas nasais.

Então, cada uma das vogais pode ser reproduzida tanto por simples ressonância da boca como por dupla ressonância, que atravessa parte da coluna de ar das fossas nasais, quer dizer, a cada uma das cinco vogais orais corresponde outra nasal (ou nasalizada). A nasalização requer menos esforço para as vogais fechadas do que para as abertas, e no Brasil ocorre somente a série de vogais fechadas /ã/, /ẽ/, /ĩ/, /õ/, /ũ/ (SAID ALI, 1964). A vogal baixa, portanto, aberta /a/, quando nasalizada, fecha-se, e realiza-se foneticamente como um chuá nasalizado.

Há pelo menos duas correntes teóricas para a interpretação das vogais nasalizadas. Uma delas defende a existência das vogais nasalizadas, e é essa concepção que se adota aqui. A outra advoga a existência de um elemento nasal acrescido à vogal, ou seja, trata-se de um grupo de dois fonemas que se combinam na sílaba: vogal e elemento nasal, mas é preciso encontrar um traço que caracterize as vogais nasais em termos fonêmicos, e esse traço deve-se procurar na constituição silábica (CÂMARA JR., 1986).

A língua portuguesa teria a nasalidade fonológica e a nasalidade que ocorre por assimilação à consoante nasal da sílaba seguinte. Nesse segundo caso, em contato com uma consoante nasal (“m” e “n”) da sílaba seguinte, uma vogal ficaria levemente nasalizada (“rima”, “comeu”). Nessa posição, as vogais da língua portuguesa sofrem redução (CÂMARA JR., 1986; 1997). Esse caso é possível sem nasalização, por isso, não é distintivo, ou seja, ela não é funcionalmente válida. Essa circunstância é mais perceptível na pronúncia da letra “a”, como em “cama” → /kˈãma/. Soam como nasais as vogais seguidas principalmente de “nh” (BECHARA, 1973).

É fonológica a nasalidade da vogal transmitida por uma consoante nasal na mesma sílaba, como em “manso” → /mˈãsu/, em oposição à nasalidade não fonológica, que resulta do contato com uma consoante nasal da sílaba seguinte, como mencionado. Assim, vogal nasal é a seqüência vogal+consoante nasal, na mesma sílaba, de modo que a nasalização da vogal é conseqüência obrigatória em português do travamento da sílaba por uma consoante nasal pós-vocálica (CÂMARA JR., 1986).

A nasalidade pura da vogal não existe fonologicamente, porque por meio dela não se cria contraste distintivo com a vogal travada por consoante nasal (CÂMARA JR., 1986). Então, “ã” equivale a “an” → /ã/, assim como “õ” equivale a “on” e “om” → /õ/. A ortografia da língua portuguesa marca graficamente apenas a nasalidade dessas duas vogais, em casos em que não as sucede consoante nasal. Diante disso, as cinco vogais do português oficial podem figurar como nasais ou nasalizadas.

A nasalização vocálica não é em si um fato fonológico em português, de forma que o que garante a ela essa característica (fonológica) não deve ser a mera nasalização da vogal. Trata-se da circunstância de haver aí uma vogal travada por um elemento consonântico nasal, como exposto. Nessa perspectiva, o ditongo nasal também é analisado como ditongo mais elemento nasal (CÂMARA JR., 1997).

A corrente que defende a existência das vogais nasalizadas ou nasais parte do pressuposto de que o elemento consonantal nasal que as sucede não é um fonema, e sim um fenômeno de antecipação da consoante que o sucede, iniciando a sílaba seguinte, portanto, é por ela condicionado e decorre da co-articulação. Há dois argumentos para refutar o elemento consonantal nasal como fonema: 1) em virtude de

esse segmento ser condicionado pela consoante seguinte, ele pode resultar numa consoante velar nasal [ɴ], se tal consoante seguinte for /k/ ou /g/, como em “ronca” → /Ɽˈõka/ e “sangue” → /sˈãgi/. Ora, sabe-se que [ɴ] não é fonema no português, pois se trata de condicionamento provocado pela consoante seguinte; 2) não é possível encontrar par mínimo em que tais elementos consonantais sejam intercambiáveis no mesmo contexto fonético, pelos mesmos argumentos já expostos.

A nasalidade vocálica em português é uma questão polêmica.

1.4.3 Semivogal

As controversas, mas assim consideradas, semivogais da língua portuguesa, representadas na grafia por “e”, “i”, “o”, “u”, “m” e “n” e grafofonemicamente por /w/ e /y/ (ou /j/), definem-se na história, com base em Almeida (1999), da forma que segue.

Semivogais chamavam-se as vogais /i/ e /u/ em situações em que partilhavam da natureza a um tempo das vogais e das consoantes. De acordo com a posição que ocupavam, conservavam-se imutáveis ou se transformavam na grafia, passando o “i” a “j”, o “u” a “v”, o que não quer dizer que todo o “j” correspondesse a /i/, que todo o “v” correspondesse a /u/. Os escritos antigos exigem cuidado de leitura, pois neles a letra “i” representa também o som /ʒ/ e a letra “u” representa também o som /v/: *iá* (“já” → /ʒˈa/), *ouui* (“ouvi” → /oʍvˈi/).

Para o português de hoje, semivogal pode ser considerada designação de sentido fonético-histórico, sem muita utilidade prática. No entanto, ainda é válida em outras línguas, como na inglesa, em que as letras “w” e “y” participam da natureza de vogal e de consoante. Diante dessa concepção, pode-se sustentar que o “m” do português, quando finaliza vocábulos, é semivogal, pois ora assim mesmo se grafa (“reprovam” → /Ɽeprˈɔvãw/) ora se grafa “o” (“Cristóvão” → /kriStˈɔvãw/) o mesmo som /w/, que então tem dupla representação gráfica, lá por uma letra consoante, cá por uma letra vogal.

Nesse sentido, o alfabeto utilizado em português veio do latim com duas modificações:

1 – O *u* e o *v* em latim – e no português antigo – escreviam-se da mesma maneira, em forma de *v*. Por isso se vêem em fachadas de prédios ou em escritos de importância inscrições como *TEATRO MVNICIPAL*, *CVRIA METROPOLITANA*.

2 – Igualmente, o *i* e o *j* confundiam-se graficamente na única forma I (*i*). Por isso se vêem encabeçando as imagens do crucifixo as iniciais INRI, cujos *ii* correspondem a *jj: I*(J)*esus Nazarenus Rex I*(J)*udaeorum*.

No próprio latim já os gramáticos distinguiam na pronúncia o *v* vogal (= u) do *v* consoante, mas distinção gráfica o português só começou a fazer do século XVI em diante, e até boa parte do XIX ainda se escrevia *dvvida* (= "dúvida").[19]

Ao tratar das vogais assilábicas, Câmara Jr. (1986, p.46) discute uma questão pertinente à possibilidade de se encontrar uma consoante /r/ brando depois de ditongo:

> [...] Com efeito essa consoante [o /r/ brando] só existe em português [em alguns encontros consonantais, como "cr" e "br", e] entre vogais. Aí cria uma oposição com /r/ forte, como vimos nos pares *era*:*erra*, *caro*:*carro*, *foro*:*forro*, *coro*:*corro* e assim por diante. Já entre consoante e vogal, como em posição inicial, só há /r/ forte (*guelra*, *Israel*, como *rato* etc.). Em face dessa propriedade fonêmica do /r/ fraco, a sua presença entre ditongo e vogal, como em *Laura*, *eira*, *europeu* e assim por diante, nos força a interpretar a vogal assilábica, mesmo em termos fonêmicos, como vogal (alofone assilábico de uma vogal) e nunca como uma consoante.

Para esse autor, as semivogais são vogais reduzidas assilábicas, e não consoantes. A leitura da semivogal não varia, é como a das vogais "i" e "u", segundo a intuição dos falantes. Então, as semivogais partilham as propriedades de /i/ e /u/, mesmo porque não há obstáculo à passagem da corrente de ar em sua emissão, diferentemente de o que ocorre com as consoantes.

As semivogais não podem ser centro de sílaba, portanto, não podem ser tônicas. São sons sempre reduzidos. Elas ocupam uma posição dita assilábica, ou seja, em vez de ser o centro da sílaba, como as vogais, ficam em suas margens, como as consoantes. O resultado disso é uma vogal modificada por outra na mesma sílaba. Isso constitui o

[19] Essa ambigüidade está presente no nome da letra dáblio que, em inglês, é *double u* ("u" duplo) e, em francês, *double v* ("v" duplo), por exemplo.

ditongo (CÂMARA JR., 1986).[20] Outras correntes definem ditongo como o encontro de uma vogal com uma semivogal na mesma sílaba. As semivogais não iniciam palavras em português, nem sílabas internas, mesmo porque sua forte tendência é estar após a vogal, e não antes dela.

Diante dessa exposição, um quadro possível das vogais do português do Brasil, segundo o timbre, apresenta-se a seguir.

Quadro 1: Timbres das vogais do PB de acordo com sua posição na palavra

Vogal	**Posição**	**Timbre**	**Exemplo**
a	tônica átona ambas	aberto reduzido (nasal)	será, carro bateu ande, órfã
e	ambas átona ambas	fechado reduzido (nasal)	preta, pegou saudade lenço, lenhoso
ɛ	tônica, subtônica* átona	aberto	bela pezinho eticamente
i	tônica átona ambas	fechado reduzido (nasal)	caí pisar cainha, inhame
o	ambas átona ambas	fechado reduzido (nasal)	bolsa, bordel branco fronha, ondular
ɔ	tônica, subtônica* átona	aberto	agora sozinho otimamente
u	tônica átona ambas	fechado reduzido (nasal)	saúva fábula unha, afundar
* na maioria das variedades sociolingüísticas			

Fonte: Adaptado de Bechara (1973), Said Ali (1964) e NGB (1959).

[20] Em outra obra, Câmara Jr. (1997) discute o papel das semivogais e sugere considerá-las consoantes, mas, em sua última obra, *Estrutura da língua portuguesa* (1970), a qual pretendia ser uma gramática, mas ficou inacabada, dada a morte do autor, ele as considera semivogais, vogais assilábicas.

Timbre é o efeito acústico resultante das diversas conformações da cavidade bucal, como, por exemplo, a distância entre a língua e o céu da boca, que é máxima para o /a/, a mais aberta das vogais. O timbre é determinado pelos movimentos do maxilar inferior e da língua no sentido horizontal e vertical, além da forma dos lábios, arredondados ou distensos. O timbre aberto pode ser designado como baixo e o fechado, como alto. O timbre é o traço distintivo das vogais (BECHARA, 1973). Alguns autores consideram que as vogais escritas acentuadas "í" e "ú" têm timbre agudo na pronúncia. O /a/ fechado existe em Portugal, mas não no Brasil.

1.4.4 Agrupamentos de Vogais

Tal como ocorre com as consoantes, também há grupos de vogais nos vocábulos. Controversamente, esses grupos podem ser ditongos, tritongos e hiatos, conforme o número de vogais que encerram e conforme a pronúncia a que obedecem. A corrente teórica que define ditongo como o encontro de uma vogal e de uma semivogal na mesma sílaba não colocaria ditongos e tritongos como encontro entre vogais.

A aceitação dos ditongos pela Nomenclatura Brasileira (NGB, 1959) provoca incoerências e confusão na língua, segundo vários autores. Um problema singularmente sério para a descrição da estrutura silábica em português é decidir se realmente há ditongos em português (CÂMARA JR., 1986).

Ditongo seria o encontro de uma vogal com uma semivogal, a combinação de duas vogais pronunciadas uma com força e clareza a outra, fracamente – por isso chamada semivogal –, e ambas pertencentes à mesma sílaba (SAID ALI, 1964). No ditongo, verifica-se neutralização intensa, e uma vogal é silábica e outra é assilábica (CÂMARA JR., 1986). Já, no hiato, as duas são silábicas contíguas ("álc*oo*l") (*Idem*). A diferença está na vogal tônica seguida de vogal átona ("sai") e vogal átona seguida de vogal tônica ("saí"). Em regra, o ditongo não é um traço fonêmico geral do português (CÂMARA JR., 1986).

Em certos casos de ditongos, em lugar de “i” e “u”, pode-se grafar a semivogal “e” e “o”, respectivamente, em observância às convenções do sistema ortográfico vigente (BECHARA, 1973). Via de regra, o ditongo pode ser crescente ou decrescente. Ele é dito decrescente quando a abertura da cavidade bucal decresce (vogal+semivogal), e pode ser oral ou nasal. É crescente (semivogal+vogal) quando a cavidade bucal começa do fechamento para a abertura. O chamado ditongo crescente é mais comum em final de palavra.

Se há ditongos em português, os verdadeiros são os decrescentes (CÂMARA JR., 1997; BISOL, 1989; SAID ALI, 1964), como em “afoito” → /afˈoytu/, posto que os crescentes podem se realizar como hiato (“miado” → [miˈadu], [myˈadu]). A distinção entre ditongo crescente e hiato anula-se muitas vezes em português. Os chamados ditongos crescentes observam-se na pronúncia lusitana e foram descritos por Gonçalves Viana, mas na pronúncia brasileira são indecisos e variáveis (SAID ALI, 1964).

O grupo “qu”+vogal “a” ou “o” é considerado também ditongo crescente, esse sim inseparável, como em “quase” → /kwˈazi/. Outro ditongo crescente que não se separa é o grupo “gu”+“a”, como em “água”, “aguar” → /ˈagwa/, /agwˈaR/. O trema também torna inseparável o grupo “q-gü”+“e” ou “i”, como em “agüentar”, “sagüi”, “freqüentar” → /agwˈẽtar/, /sagwˈi/, /frekwẽtˈaR/. São circunstâncias excepcionais, em que esses ditongos são perfeitos. O grupo “guo” separa-se em alguns casos, como em “averiguo” → /averigˈuu/. A explanação de vários autores permite considerar hiatos a maioria dos ditongos crescentes.

Na escrita, os ditongos crescentes oficiais da língua portuguesa são 11 (NGB, 1959): “ea” (“ígnea”), “eo” (“marmóreo”), “ia” (“história”), “ie” (“superfície”), “io” (“repertório”), “oa” (“mágoa”), “oe” (“perdoe”), “ua” (“decídua”), “ue” (“acentue”), “ui” (“contribuí”), “uo” (“árduo”). Já os ditongos decrescentes são oito orais: “ai” (“saibro”), “au” (“aura”), “éi” – “ei” (“pastéis” – “aceite”), “éu” – “eu” (“fogaréu” – “europeu”), “iu” (“ouviu”), “ói” – “oi” (“corrói”, “coitado”), “ou” (“couro”), “ui” (“circuito”); e cinco nasais: “ãe” (“mamãe”), “ão” – “am” (“bênção” – “correram”),

“em” (“convém”) pronunciado /ẽy/, “õe” (“corações”), “ui” (“muito”) pronunciado /uỹ/.

Tem-se o tritongo quando uma vogal forte se acha entre duas fracas e as três fazem parte da mesma sílaba (SAID ALI, 1964). Então, na escrita, tritongo é o grupo constituído por uma vogal acentuada, ladeada por duas outras (“ag*üei*”, “q*uai*s”, “sag*uão*”), o que seria semivogal+vogal+semivogal. O tritongo aparece na transcrição fonológica: /agwˈey/, /kwˈayS/, /sagwˈãw/, quando também a pronúncia mostra que a vogal “o” do tritongo nasal “uão” transforma-se na semivogal /w/. Nessa condição, também há o tritongo /waw/: “igual” → /igwˈaw/.

Como os ditongos, os tritongos podem ser orais ou nasais. O ditongo nasal é analisado como ditongo mais arquifonema nasal (CÂMARA JR., 1997). Uma das características distintivas da língua portuguesa é a nasalização que ocorre em alguns vocábulos.

A diferença entre as pronúncias brasileira e lusitana provoca incompatibilidades nas definições da NGB – que tenta igualar os dois sistemas – e oscilações de opinião entre teóricos brasileiros. A pronúncia lusitana evidencia ditongos crescentes (como em “poema” → [pwˈema]) e tritongos (como em “fiéis” → [fyˈɛyS]) onde no Brasil há hiato /poˈema/ e ditongo /fiˈɛyS/. A pronúncia brasileira tende a apoiar-se na primeira das vogais, de forma a decompor o tritongo em duas sílabas (SAID ALI, 1964), vogal+ditongo, portanto. A partir disso, restam como tritongos os agrupamentos vocálicos que antecedem consoante velar, como nos exemplos anteriores. Casos como “saudade”, que a NGB diz não ser ditongo, são controvertidos, uma vez que se pronuncia o “au” de “saudade” da mesma forma que o “al” de “saldado” → /sawdˈadi/, /sawdˈadu/, na maioria das variedades sociolingüísticas do Brasil.

Apesar da contenda, permanece essa nomenclatura nas gramáticas escolares e nos textos de lingüística. Ao sistematizar eletronicamente a língua, uma posição fixa e lógica deve ser tomada em todos esses casos, pois há necessidade de se estabelecer um padrão.

1.4.5 Harmonia Vocálica e Vogais Abertas Não Acentuadas

Harmonia vocálica é a influência que o som de uma vogal, numa palavra, exerce sobre outros sons, vizinhos ou próximos, e que os torna mais semelhantes e simétricos.

Em algumas variedades sociolingüísticas, as oposições entre /o/ e /u/ em posição penúltima átona ficam prejudicadas pela tendência de harmonizar a altura da vogal pretônica com a da vogal tônica. A rigor, diante de /i/ e /u/ tônicos, /e/ e /o/ aparecem com firmeza em vocábulos inusitados na linguagem corriqueira. Coloquialmente, essa distinção é praticamente gráfica, como em "comprido" e "cumprido", porque, correntemente, na maioria das variedades sociolingüísticas, pronunciam-se /kũprˈidu/ os dois vocábulos (CÂMARA JR., 1986). No entanto, falares do Brasil preservam essa distinção, como no Sul, em que há, coloquialmente, "cozinho" → [kozˈĩɲu], "amendoim" → [amẽdoˈi]/. Em se tratando de transposição de /e/ para /i/, pode haver várias opções, como ocorre com "testemunho" → [teStemˈũɲu], [teStimˈũɲu] e [tiStimˈũɲu].

Configurem casos de harmonia vocálica ou não, as vogais escritas "e" e "o", quando descodificadas como as vogais abertas, mas não marcadas graficamente, causam dificuldade de leitura, pois muitas vezes não há como prever essa pronúncia. Cabe expor alguns casos desses.

Em palavras terminadas em "osa", esse "o" é aberto e tônico (regra das paroxítonas terminadas em "a") na descodificação: "generosa" → /ʒenerˈɔza/, "rosa" → /Ɽˈɔza/, mesmo que a palavra esteja no plural. Isso ocorre também com plurais masculinos dessas mesmas palavras: "generosos" → /ʒenerˈɔsuS/, "ansiosos" → /ãsiˈɔsuS/. Ainda, ocorre no plural de algumas palavras paroxítonas terminadas em sílaba aberta em "o": "porcos" → /pˈɔRkuS/, "corvos" → /kˈɔRvuS/.

Esse fenômeno é observado em algumas formas verbais, quando então as letras "o" e "e" podem corresponder aos fonemas /ɛ / e /ɔ/, como em "disseste", "espera", "comemoro", "remove" → /disˈɛSti/, /eSpˈɛra/, /komemˈɔru/, /Ɽemˈɔvi/. É uma

sistematização complexa e em nível morfofonêmico. Certamente, não é uma distinção fácil de ensinar na escola, mas o uso torna-a óbvia aos falantes nativos.

Na maioria das situações, a descodificação correta do timbre aberto representado pelas letras “e” e “o”, quando não houver acento gráfico, depende da aplicação de conhecimentos morfossintáticos e semânticos à posição que o item ocupa na frase, combinados com o emparelhamento no léxico mental ortográfico e a respectiva realização de seu valor fonológico. Há casos que somente a gramática pode resolver, quando então o contexto maior diz se se trata de verbo, como “gosto” (ó), ou substantivo, como “gosto” (ô) (SCLIAR-CABRAL, 2003), por exemplo. Em muitos vocábulos, há dúvida quanto ao timbre das vogais (BECHARA, 1973).

1.4.6 Consoante

As vogais produzem-se mediante modificações na boca e as consoantes são produto da interrupção da correnteza de ar expelida pelos pulmões. Enquanto as vogais produzem-se livremente, as consoantes produzem-se rompendo um obstáculo maior ou menor à passagem do ar expelido pelos pulmões. No caso das oclusivas, ao se desobstruir a passagem do ar, ouve-se o som da vogal modificada conforme a consoante oclusiva que a precedeu (fenômeno da co-articulação). No caso das demais consoantes, ouve-se o ruído proveniente da vibração das moléculas de ar entre as partes que impediam a saída do ar e, logo a seguir, o som de uma vogal, resultante dessa desobstrução. Daí vem o nome consoante (com+soante), ou seja, som acompanhado de vogal.

Então, consoante é um som produzido com obstrução total, constritiva ou parcial à passagem da corrente de ar. Considerando-se as oposições distintivas, tal obstrução pode ser oclusão ou fechamento, constrição ou aperto, oclusão parcial que desvia a direção da corrente de ar ou tremulação da língua, que provoca vibração à corrente de ar. Isso dá para as consoantes as seguintes ordens: oclusivas (auditivamente plosivas); constritivas (auditivamente fricativas); nasais com oclusão ou constrição na boca, mas com ressonância plena nas fossas nasais e comunicação entre boca e nariz;

laterais, com oclusão num ponto do centro da língua e desvio lateral da corrente de ar; vibrantes, com a vibração rápida ou prolongada da língua, ou da úvula, ou fricção faríngea (CÂMARA JR., 1986). As letras que representam as consoantes que compõem o sistema ortográfico oficial do português são 19: “b”, “c”, “ç”, “d”, “f”, “g”, “h”, “j”, “l”, “m”, “n”, “p”, “q”, “r”, “s”, “t”, “v”, “x” e “z”.

A posição mais favorável da consoante em português é antes da vogal da sílaba. Pode ser intervocálica, quando separa duas sílabas, ou não intervocálica, quando inicia vocábulo e quando for medial, e vier depois de outra consoante da sílaba precedente (CÂMARA JR., 1986). Ainda, na posição não intervocálica, pode vir depois de outra consoante na mesma sílaba.

As consoantes intervocálicas em português apresentam articulação um tanto enfraquecida pelo ambiente vocálico em que se acham. São por isso alofones posicionais das consoantes não intervocálicas correspondentes, de articulação mais firme. Em compensação, certas consoantes faltam em posição não intervocálica (CÂMARA JR., 1986). Na escrita, as consoantes que travam sílaba em português são normalmente “m”, “n”, “r”, “l”, “s”, “z”, mas pode haver casos de sílaba travada por “b” (“absorver”), “c” (“bactéria”), “d” (“advertência”), “f” (“afta”), “g” (diagnóstico), “p” (“apto”), nesses casos, geralmente na pronúncia insere-se um /i/ reduzido após a consoante que trava a sílaba, uma vez que essa combinação não é comum ao português.

Nesse sentido, se a sílaba aberta e a seqüência consoante+vogal (CV) são o padrão da sílaba em português, todas as consoantes podem se agrupar com uma vogal. Quando a sílaba é CCV, somente as laterais e vibrantes anteriores assumem a posição pré-vocálica – /l/ e /r/ – como em “flor” e “crer”. Com outros grupos ocorre epêntese de uma vogal. Se a sílaba é CVC, na posição pós-vocálica, as consoantes possíveis são as que travam sílaba, já mencionadas, embora “m” e “n” nessa posição não se realizem como fonemas, e sim como marca de nasalização da vogal precedente. Nas sílabas travadas por “l” na escrita, essa letra, ao ser descodificada como fonema, sofre uma mutação chamada vocalização da consoante (CÂMARA JR., 1986) e, então, representa a semivogal /w/, como em “mel” → /mɛw/. Assim desaparece da língua a

possibilidade de haver sílaba travada por /l/ (CÂMARA JR., 1986), que se descodifica como /u/ assilábico, /w/, portanto. Todavia, o "l" velar [ɬ] aparece, por exemplo, em regiões do Rio Grande do Sul.

A consoante "x" merece comentários à parte. A maior dificuldade ortográfica e, conseqüentemente, grafofonêmica da língua portuguesa é o uso do "x", porque seu uso no português não obedece a regras sólidas. É o grafema cuja descodificação é a mais imprevisível da língua portuguesa.

A letra "x" representa cinco sons em português:

1) som alfabético (padrão), ou seja, chiante /ʃ/, que era desconhecido dos romanos, pois não havia som chiado no latim: "xeque", "praxe", "vexar" → /ʃˈɛki/, /prˈaʃi/, /veʃˈaR/;
2) som sibilante surdo /s/: "sintaxe", "trouxe" → /sĩtˈasi/, /trˈowsi/;
3) som sibilante sonoro /z/: "exame", "existir", "execrar", "exangue" → /ezˈãmi/, /eziStˈiR/, /ezekrˈaR/, /ezˈãgi/;
4) som de /kis/ ou /kiS/: "látex", "nexo", "complexo" → /lˈatekiS/, /nˈɛkisu/, /kõplˈɛkisu/;
5) realizações do arquifonema |S|: "texto", "explicar" → /tˈeStu/, /eSplikˈaR/.

A grafia de "x" provém:

a) de um "x" ou "xs": "enxugar" (de *exucare* ou *exsucare*);

b) da combinação "sc": "mexer" (de *miscere*), "faixa" (de *fasciam*);

c) de um "s": "bexiga" (de *vesicam*), "enxertar" (de *insertare*), "puxar" (de *pulsare*);

d) de dois "s": "graxo" (de *grassum*), "roxo" (de *russeum*);

e) do dígrafo inglês "sh": "xampu", "xerife".

Também merece destaque a consoante "s". Na maioria das vezes, o "s" das palavras em português corresponde à letra de origem: "vaso" – *vasum*, "peso" – *pensum*. O

étimo da palavra faz conservar seu "s", então, não se justificam grafias como *portuguez* e *apezar*. A letra "s" pode representar dois sons: sibilante surdo e sibilante sonoro. Seu som alfabético (padrão) é sibilante surdo /s/ ("salgar" → /sawgˈaR/) e o som sibilante sonoro é som acidental (não padrão), que corresponde a sua leitura como /z/. Por exemplo, "obséquio" tem seu som acidental, ou seja, /z/: /obizˈɛkiw/, ao passo que "observar" → /obiseRvˈaR/ tem seu som alfabético. Há casos em que somente o uso consagrado determina o som que a letra "s" representa. Ainda, em se tratando de prefixo terminado em "s", na leitura, essa letra terá som sonoro quando a ela seguir vogal e terá som surdo quando o elemento posposto ao morfema ou prefixo tiver "s" inicial etimológico: "transubstanciação" → /trãsubiStãsiasˈãw/, "transabdominal" → /trãzabidominˈaw/. Na realidade, trata-se de "ss", um do prefixo e outro da palavra que o recebe – etimológicos, portanto –, e o último desapareceu por causa do "s" que já existe no prefixo: trans+substanciação. Em latim, o "c" era pronunciado como /k/.

1.4.7 Agrupamentos de Consoantes

Na língua escrita, há grupos de letras que representam consoantes não intercaladas por vogais. Essas junções podem ser dígrafos, encontros consonantais perfeitos e encontros consonantais imperfeitos. Dígrafo é a junção de duas letras para representar um único fonema. Os dígrafos da língua portuguesa são: "ch", "lh", "nh", "ss", "rr", "qu", "gu", "xc", "sc"[21] que correspondem aos fonemas /ʃ, ʎ, ɲ, s, ʀ̥, k, g, s, s/. Se for aceito o papel das letras "m" e "n" em final de sílaba para nasalizar, todas as letras que representam vogal seguida daquelas letras na mesma

[21] Em latim, não existiam "ch", "lh" e "nh". O "ch" corresponde ao som alfabético que a letra "x" representa, ou seja, /ʃ/, e pode ter diversas origens. O "lh" representa, quanto ao som, o duplo /l/ (/l/ molhado) do espanhol (*manilla*) ou o /l/ latino, que tem por função evitar hiatos (mul*i*erem → "mu*lh*er", fol*i*a → "fo*lh*a"). O "nh", a exemplo do "lh", corresponde ao /n/ com a mesma finalidade: se*n*iorem – "se*nh*or", ve*n*io – "ve*nh*o". Em regiões rurais, por exemplo, na fala, o hiato é abrandado de outra forma: "mi*lh*o" → [mˈiu], "mu*lh*er" → [muiˈɛ] (ALMEIDA, 1999). O hiato aumenta uma sílaba na palavra, portanto, é menos econômico.

sílaba também seriam dígrafos, como “em” e “an” em “tempo” e “amante”, que representam /ẽ/ e /ã/.

Os encontros consonantais perfeitos ocorrem quando as consoantes do grupo são pronunciadas e pertencem à mesma sílaba, nesse caso, geralmente a segunda consoante é “l” ou “r”: “cr”, “br”, “cl”, “pl” etc. Esse encontro (consoante+“l” ou “r”) é mais forte, ou seja, o grupo é inseparável. Nos encontros consonantais imperfeitos e até em alguns dígrafos, as letras que os compõem pertencem a sílabas diferentes: “bs” (“absorver”), “ft” (“afta”), “mp” (“campo”), “sc” (“nascer”), “xc” (“excelente”) etc., e pode aparecer uma vogal epentética na pronúncia (“advogado”, “subseção” → /adivogˈadu, subisesˈãw/). Nos dígrafos, duas letras representam um só fonema, conforme mencionado: /krˈesa/.

O português abrandou algumas consonâncias do latim, pois há casos em que uma consoante surda latina transformou-se em sua respectiva homorgânica sonora. Assim, ***c****atum* transformou-se em “**g**ato”, mediante sonorização do “c” duro inicial, ou seja, /k/ → /g/; felici*ta*tem e ae*t*atem transformaram-se em “felici*da*de” e “i*da*de”, em que /t/ → /d/; sa*p*ere transformou-se em “sa*b*er”, em que /p/ → /b/. Também o /b/ latino transforma-se freqüentemente em /v/: ama*b*am → “ama*v*a”, em que a consoante oclusiva passa a constritiva (ALMEIDA, 1999).

Um quadro da classificação oficial das letras que representam as consoantes do português e dos fonemas consonantais do português do Brasil configura-se como segue.

Esses quadros são apenas didáticos, porque são fiéis à NGB, e os critérios que nortearam o programa de conversão dos grafemas em fonemas foram fonológicos. Assim, classificações que constam nos quadros 2 e 3 referem detalhes fonéticos que não são pertinentes a uma descrição fonológica, como as distinções entre bilabiais e labiodentais, entre linguodentais e alveolares. Por outro lado, no Quadro 2, deixam de figurar dígrafos, como “qu”, “gu”, “xc”, “sc”, “xç” e “sç”, e não se faz referência ao valor que os grafemas têm, em virtude do contexto grafêmico.

Quadro 2: Classificação das letras que representam as consoantes portuguesas

Modo de articulação		Oclusivas			Constritivas					
					Fricativas		Laterais		Vibrantes	
Papel das cordas vocais		Surdas	Sonoras		Surdas	Sonoras	Surdas	Sonoras	Surdas	Sonoras
Papel das cavidades oral e nasal		Orais	Orais	Nasais	Orais	Orais	Orais	Orais	Orais	Orais
Ponto de articulação	Bilabiais	p	b	m	-	-	-	-	-	-
	Labiodentais	-	-	-	f	v	-	-	-	-
	Linguodentais	t	d	-	-	-	-	-	-	-
	Alveolares	-	-	n	s, c, ç	z, s (brando)	-	l	r (forte), rr	r (brando)
	Palatais	-	-	nh	x, ch	j, g (brando)	-	lh	-	-
	Velares	c (duro), q	G (duro)	-	-	-	-	-	-	-

Quadro 3: Classificação dos fonemas consonantais portugueses

Modo de articulação		Oclusivos			Constritivos					
					Fricativos		Laterais		Vibrantes	
Papel das cordas vocais		Surdos	Sonoros		Surdos	Sonoros	Surdos	Sonoros	Surdos	Sonoros
Papel das cavidades oral e nasal		Orais	Orais	Nasais	Orais	Orais	Orais	Orais	Orais	Orais
Ponto de articulação	Bilabiais	/p/	/b/	/m/	-	-	-	-	-	-
	Labiodentais	-	-	-	/f/	/v/	-	-	-	-
	Linguodentais	/t/	/d/	-	-	-	-	-	-	-
	Alveolares	-	-	/n/	/s/	/z/	-	/l/	/ɾ̥/	/r/
	Palatais	-	-	/ɲ/	/ʃ/	/ʒ/	-	/ʎ/	-	-
	Velares	/k/	/g/	-	-	-	-	-	-	-

Fonte: Adaptados de Bechara (1973) e Said Ali (1964).

1.5 PROSÓDIA

Acentuação é o modo de fazer sobressair um som dentre muitos no vocábulo ou na frase. Há os acentos musical (de altura), de intensidade (tônico) e de duração. O acento de intensidade resulta da expiração mais forte ao pronunciar sílabas, enquanto o acento musical ocorre na interrogação, exclamação, na linguagem emocional (SAID ALI, 1964). O acento musical resulta dos contrastes de freqüência (intervalos), em geral, das vogais: tais freqüências são os tons fundamentais produzidos na laringe. A duração resulta dos contrastes entre sons breves e longos, em geral nas vogais, como no grego e no latim. Na língua portuguesa há os acentos de intensidade e de altura, e a duração ocorre redundantemente com a intensidade.

A prosódia – cuja origem é grega, *prosoidía*, e significa acento que se põe sobre as vogais (AURÉLIO, 1999) – ou organização dos fatos do acento compreende os fenômenos distintivos que caracterizam e opõem as unidades de expressão mais longas do que o fonema isolado (sílabas, seqüências de sílabas, grupos). É uma parte da fonética e incumbe-se de estudar a tonicidade dos sons reunidos, ou seja, a pronúncia das palavras (MALMBERG, 1993).

A utilização da prosódia varia muito entre as línguas. As línguas românicas meridionais e parte das línguas germânicas distinguem palavras e formas unicamente com a ajuda da distribuição do acento na palavra – em português, vale o exemplo: “cantara” e “cantará”. Não há essa possibilidade no francês, porque nessa língua o acento está sempre na última sílaba da palavra. Em todas as línguas existem diferentes tipos de variação de acento da frase, por meio da melodia ou da entoação, por meio de diferenças de intensidade. A entoação muitas vezes serve para marcar oposições gramaticais (MALMBERG, 1993).

A tonicidade (o acento) resulta de uma força maior expiratória ou intensidade de emissão da vogal de uma sílaba em contraste com as demais vogais silábicas. Em português, ela pode incidir na última, penúltima, antepenúltima ou, mais raramente, na quarta última sílaba de um vocábulo fonológico. As sílabas pretônicas são mais fracas do que as postônicas. O vocábulo fonológico é bem delimitado em português, e sua

marca nítida é o acento (CÂMARA JR., 1986). A posição tônica dá, nítida e plenamente, os traços distintivos vocálicos (CÂMARA JR., 1997).

O acento é livre no sentido de que sua posição não depende da estrutura fonêmica do vocábulo. Não há em português terminações de fonemas que imponham dada acentuação, há maior freqüência, fonologicamente indeterminável, para dada terminação. No entanto, há um tipo de acentuação que caracteriza a língua portuguesa: o paroxítono, que confere à língua ritmo grave.[22] O português do Brasil diferencia-se do de Portugal por maior número de vocábulos oxítonos, incorporados das línguas indígenas e africanas que aqui conviveram com o português no passado (*Idem*).

1.5.1 Sílaba

A sílaba é a unidade superior, na qual os fonemas (vogais e consoantes) combinam-se para funcionar na enunciação (CÂMARA JR., 1997). Esta seção trata especificamente da sílaba, mas a sílaba acompanhou toda a teoria até agora, pois é fundamental para a descrição do funcionamento da língua.

A teoria da sílaba é o foco recente da fonologia. Antes a sílaba era estudada, mas somente a partir dos anos 1960 passou-se a considerar com mais afinco unidades maiores das palavras, ou seja, grupos de fonemas. Esses estudos valem-se dos anteriores e, a partir deles, apresentam inovações na área.

A sílaba é uma divisão espontânea e profundamente examinada pela fonologia. Seus tipos de estrutura marcam caracteristicamente as línguas. Não é o fonema, mas sim a sílaba a estrutura fonêmica elementar (JAKOBSON, 1967 *apud* CÂMARA JR., 1986).

Câmara Jr. (1997) entende a sílaba em português como um conjunto de posições a ser ocupadas por fonemas específicos:

[22] Em comparação com outras línguas latinas, o espanhol também tem ritmo grave, embora talvez mais suave do que o português (MALMBERG, 1993); o italiano, esdrúxulo; e o francês – que tem acento fixo, e é constituído por vocábulos oxítonos – tem ritmo agudo (CÂMARA JR., 1986).

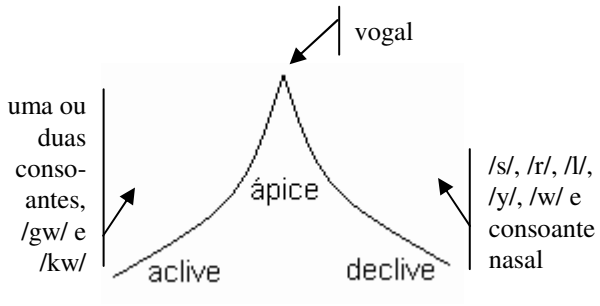


Figura 2: Esquema silábico do português

O núcleo da sílaba é a única posição indispensável em português e deve ser sempre ocupado por uma vogal, que é o som preponderante da sílaba. O aclive é basicamente ocupado pelas consoantes, então, é uma categoria que pode não estar presente na sílaba. Todas as consoantes fonológicas do português podem ocupar essa posição, mas as seqüências de duas consoantes nessa posição são restritas, e não são admitidas as seqüências /dl/ e /vr/ (esta não é admitida somente em início de vocábulo), por exemplo. Cabe comentar que não há menção a /y/ e /w/ no aclive. Já maiores restrições são feitas ao que pode estar no declive, que aceita certas consoantes, as semivogais /y/ e /w/ – as quais não são pico de sílaba e têm características tanto de vogais como de consoantes –, mas também pode estar vazio.

Na língua portuguesa predominam sílabas livres ou abertas – que são as que têm terminação silábica, ou seja, terminam em vogal. Nelas incluem-se as sílabas simples (V) e complexa aberta (CV). Sílabas travadas ou fechadas são as que terminam em consoantes (VC, CVC). Elas são muito menos freqüentes em português e há grande limitação das consoantes que podem figurar no aclive (*Idem*).

Ainda no tocante à estrutura silábica do português, há o caso dos empréstimos do latim clássico, introduzidos por via escrita, a partir do século XV (CÂMARA JR., 1986). Trata-se da junção gráfica de duas consoantes pronunciadas (plosiva ou fricativa labial seguida de plosiva, fricativa labial ou uma nasal). É o que ocorre em "na**ft**alina", "**ps**icose", "r**ap**to". Na pronúncia, há o que parece ser uma vogal intercalada entre essas consoantes, similar a /i/, que não pode ser desprezada, pois se equipara à pronúncia de vogais reduzidas postônicas ("rápido", "rapto") e pretônicas,

quando então a pronúncia do /i/ é instável, por vezes, quase inaudível. Essa epêntese ocorre por razão que já se mencionou. Uma consoante oclusiva não pode ser pronunciada isoladamente, então, no encontro de duas consoantes desse tipo, ocorre tal epêntese naturalmente, para facilitar a pronúncia, pois não há vogal para finalizar a sílaba, como é comum nesses casos com as consoantes líquidas (“cra”, “fle”).

De maneira geral, a delimitação silábica é nítida em português, mas há três casos em que é flutuante. Trata-se de três contextos de grupos de vogais em que entra, como primeira ou segunda vogal, uma vogal alta e átona (CÂMARA JR., 1997): a) /i/ ou /u/ precedido ou seguido de outra vogal átona (“saudade”, “variedade”); b) /i/ ou /u/ seguido por vogal tônica (“piano”, “viola”); e c) /i/ ou /u/ seguido por vogal átona em final de vocábulo (“índia”, “assíduo”). Foneticamente, podem-se entender esses casos como ditongos ou hiatos, em variação livre, sem oposição distintiva. Fonologicamente, entretanto, há uma fronteira silábica variável e não significativa (*Idem*).

Atualmente, a teoria da sílaba mais estudada é a métrica, que considera a sílaba como formada por camadas de constituintes mais ou menos relacionados entre si. Segundo essa teoria, uma sílaba (σ) consiste em um ataque (A) – que é o aclive – e em uma rima (R), e a rima consiste em um núcleo (Nu) – que é o ápice – e em uma coda (Co) – que é o declive. Essa sistematização é discutida em Clements (1985) e foi amadurecida por estudiosos brasileiros para a língua portuguesa. No entanto, não foi utilizada para os objetivos desta pesquisa, mesmo porque é o que Câmara Jr. apresenta, mas exposto de outra forma.

1.5.2 Acentuação Gráfica

O sistema de acentuação gráfica é referência importante para a leitura, portanto, baseia-se na escrita. Apesar de o sistema de acentuação vigente no português do Brasil parecer para muitos uma forma de complicar a escrita, ele facilita a leitura e torna a escrita mais coerente e objetiva. Muitos não entendem por que as palavras recebem acento gráfico, o que significa que não levam em conta um dos dois lados do sistema verbal: codificação e descodificação; escrita e leitura. Como se expôs na seção sobre

prosódia, a grande maioria das palavras da língua portuguesa é acentuada – ou seja, tem sílaba tônica –, mas esse acento não é marcado graficamente na maioria delas. Esta seção trata da acentuação marcada.

O princípio que rege o sistema de acentuação gráfica na língua portuguesa do Brasil é marcar o mínimo de palavras, ou seja, o princípio da economia de acentos gráficos. Para tanto, o sistema considera a quantidade de palavras existentes com as três tonicidades permitidas na língua portuguesa (oxítona, paroxítona e proparoxítona). Como a forma canônica do vocábulo português é a tonicidade estar na penúltima sílaba, o falante nativo tende a pronunciar dessa forma uma palavra que não conheça, intuitivamente. No entanto, palavras estrangeiras, como nomes e sobrenomes (italianos, poloneses, alemães), podem desviar-se dessa ordem.

As palavras do português do Brasil são acentuadas graficamente conforme a freqüência (que implica quantidade) de sua tonicidade na língua. Palavras cuja tonicidade é rara são acentuadas graficamente, e pequena parte das palavras cuja tonicidade é freqüente é acentuada. Dessa forma, todas as proparoxítonas são marcadas, por serem as mais raras, e poucas paroxítonas são acentuadas graficamente, por serem as mais comuns em português. Ainda, boa parte das oxítonas é acentuada, por ocuparem o segundo lugar na freqüência da língua portuguesa.

Na escrita, o acento gráfico (agudo e circunflexo) é uma poderosa marca distintiva para os usuários de uma língua, pois assinala a sílaba tônica de uma palavra e guia a leitura. O sinal diacrítico til, que nasala uma vogal, também marca tonicidade, quando ocorrer em fim de palavra que não tiver acento agudo ou circunflexo. Quando não houver acento gráfico, resta a metalinguagem, para se saber qual é a sílaba tônica.

A terminação da palavra, em seu modo escrito, é o critério que define a acentuação gráfica, à exceção de raros hiatos tônicos e acentos diferenciais. A tonicidade não marcada na grafia depende exclusivamente da terminação do vocábulo. Letras não pronunciadas não podem receber acento gráfico. No português escrito, as palavras terminam em letras que representam vogal oral e nasal, ditongo oral e nasal e na letra “n”, seguidos ou não de “s”; e em “m”; “n”; “r”; “x”; “l”; “z”; “ps”.

Os vocábulos mais numerosos da língua são os paroxítonos que terminam com as letras “e”, “o”, “a”, que representam vogais orais, seguidas ou não de “s”, então, não se acentuam palavras desse tipo. No português do Brasil, na maior parte das variedades sociolingüísticas, as vogais átonas finais grafadas com “e” e “o” neutralizam-se em favor de /i/ e /u/. Essa neutralização dá-se em função da tonicidade da palavra, pois, uma vez que a antepenúltima ou a penúltima sílaba é mais intensa, reduz-se a última sílaba, de forma que “e” e “o” reduzidos descodificam-se como /i/ e /u/. O restante das paroxítonas recebe acento gráfico para guiar a leitura.

A vogal “a” nasal, se estiver em final de vocábulo, torna a palavra oxítona (“maçã”, “talismã”, “suã”). As letras “i” e “u” quando figuram em final de palavra puxam a tonicidade para a sílaba final. Assim, a tendência de um falante nativo ao ler uma palavra desse tipo, mesmo que não a conheça, é acentuá-la na última sílaba.

Quanto à terminação “m”, são paroxítonas as palavras que terminam em “am” (“amam”) e “em” (“comem”) e oxítonas as que terminam em “im” (“curumim”), “om” (“batom”, “edredom”) e “um” (“algum”). Espera-se que, a exemplo de “am” e “em”, “om” figure em final de palavras paroxítonas, no entanto, essa terminação indica vocábulo oxítono.

Os ditongos, quando ocorrem em final de palavra, também puxam a tonicidade para a última sílaba (“plebeu”, “apertei”), então, como a semivogal, que é a última letra, não recebe acento, o acento recai na vogal do ditongo. Em final de palavra, esses ditongos são escassos, “oi”, por exemplo, é raríssimo. Mais comuns são os ditongos abertos “éi”, “ói” e “éu” que, em contraposição aos fechados, são acentuados, em respeito a sua pronúncia (“meia”, “idéia”; “boi”, “dói”; “jubileu”, “mausoléu”). Logo, trata-se de acento diferencial. Essas terminações também são raras e ocorrem mais nos plurais de alguns vocábulos (“lençóis”, “anéis”). Somente no plural há “éi” em sílaba final de vocábulo. Vocábulos que terminam em “oo”, seguido ou não de “s”, e “eem” (conjugação de alguns verbos na terceira pessoa do plural do presente do indicativo) recebem acento na primeira dessas vogais, para indicar que se trata de hiato.

Se o ditongo decrescente considera-se uma única sílaba, o chamado ditongo crescente, exceto os casos vistos em que não há separação possível, constitui duas

sílabas, portanto, quando o ditongo crescente (que na verdade não é ditongo) que finaliza palavra não é acentuado, com certeza o acento recai na sílaba que o antecede, ou seja, na antepenúltima sílaba, de forma que a palavra é proparoxítona. Como a ortografia oficial pede que se acentuem as proparoxítonas, a sílaba que antecede o ditongo crescente átono deve ser acentuada (“glória”, “ciência”).

As vogais “i” e “u”, quando estiverem sozinhas na sílaba ou seguidas da letra “s”, recebem acento agudo, se seguirem outra vogal que não forma com elas ditongo, justamente para sinalizar que a palavra tem uma sílaba a mais, formada por uma vogal (“saí”, “sai”; “saúde”, “sauna”; “balaústre”, “holocausto”). Nesse caso, há uma exceção feita à vogal “i” seguida por “nh”. Ao que parece, essa exceção deve-se a ser intuitivo pronunciar vogal+inh* como duas sílabas. Se fosse ditongo, a semivogal desapareceria na pronúncia e anular-se-ia o contraste entre “bainha” e “banha”, por exemplo. Nesse caso, há palavras com “ainh”, “oinh” e “uinh” (“cainha”, “moinho”, “fuinha”) e palavras diminutivas com “einh”, como “meinha” e “veinha”, mas não são muitas. O dicionário Aurélio (2004) não registra palavras com a combinação vogal+unh*. Assim, pela lógica do sistema de acentuação gráfica do português, caberia acento nessa circunstância, se houvesse risco de o falante ler o agrupamento de vogais como ditongo, e não como hiato.

As terminações “r”, “x”, “z” e “l” fortificam a última sílaba dos vocábulos, de forma que palavras com essas terminações são oxítonas (“amor”, “poder”; “unissex”, “genetrix”; “avestruz”, “capataz”; “varonil”, “ramal”).

Terminações em “n” não são usuais na língua portuguesa, em que, nesse caso, natural seria a letra “m”. Desse modo, a grande maioria das palavras com essa terminação é estrangeira aportuguesada e recebe acento gráfico, se não o recebe, é oxítona (“sêmen”, “xiamen”; “pídgin”, “sarin”, “elétron”, “frisson”, “tabun”). Mais comum é “ns”, que forma plural de palavras cuja forma singular termina em “m” e também das que terminam em “n”.

Se as paroxítonas terminadas em “a”, “e” e “o” não são acentuadas, então, as oxítonas terminadas assim são acentuadas, de forma a guiar o leitor e evitar erro de descodificação. Da mesma forma, as palavras terminadas em “i” e “u”, as mais

freqüentes dentre as oxítonas, se não forem oxítonas, recebem acento. É uma regra em contrário com as paroxítonas. Então, se a tendência é ler a palavra de certo modo, mas a pronúncia de uma dada palavra desviar-se disso, essa palavra será acentuada graficamente.

Os monossílabos tônicos terminados em “a”, “e” e “o”, seguidos ou não de “s” recebem acento (“pá”, “crê”, “dó”). Os nomes de algumas letras são monossílabos tônicos, então, recebem acento gráfico: “bê”, “cê”, “dê”, “ê”, “gê”, “pê”, “quê”, “tê”, “vê”, “zê”.

1.5.2.1 Acento diferencial

Persistem em português casos de acentos diferenciais. São acentos gráficos, agudo ou circunflexo, que se põem sobre as vogais “a”, “e” e “o” em alguns vocábulos tônicos para distingui-los de seus homônimos átonos ou para discernir as vogais fechadas /e/ e /o/ das respectivas abertas. Os casos mais comuns são: o acento agudo em “pára” (verbo), em contraposição a “para” (preposição); “pêlo” (substantivo) e “pélo” (verbo), em contraposição a “pelo” (combinação de preposição e artigo); “pôr” (verbo), em contraposição a “por” (preposição), “pólo” (eixo) e “pôlo” (gavião jovem); “pêra” (fruta), “pera” (preposição antiga: “para”) e “péra” (“pedra”).

O acento (circunflexo) que se usa sobre o “o” fechado de alguns vocábulos, como em “pôde” (3ª pessoa do singular do pretérito perfeito do modo indicativo do verbo “poder”), para distingui-lo de “pode” (3ª pessoa do singular do presente do indicativo desse mesmo verbo), e “fôrma” (recipiente oco) para distingui-lo de “forma” (feitio), também é diferencial. Finalmente, tal acento ocorre também com os verbos “ter”, “vir” e seus derivados, para diferenciar a terceira pessoa do singular do tempo presente do modo indicativo (ele) “tem”, “vem” e a terceira pessoa do plural do mesmo tempo e modo (eles) “detêm”, “sobrevêm”.

1.5.2.2 Trema

O sinal diacrítico trema (¨) ocorre sobreposto à vogal “u”, entre “q” ou “g” e “e” ou “i” e serve para indicar que esse “u” é pronunciado, uma vez que pode ocorrer nessa mesma posição, e não ser pronunciado (“sagüi”, “freqüente”, “bilíngüe”). Se o “u” for tônico, o que somente ocorre com poucos verbos, então recebe acento agudo (“apazigúe”), para marcar o hiato.

Segundo a teoria oficial (NGB, 1959), o trema marca a divisão do ditongo em duas sílabas, ou seja, serve para indicar que a vogal “u” não forma ditongo com a que lhe está próxima, porque a vogal “u” é pronunciada.

> Prevalece ainda o trema no “u” sempre que pronunciado depois de *g* ou *q* seguido de “e” ou “i”: *agüentar*, *argüição*, *eloqüente*, *tranqüilo*, *cinqüenta*... – É o que está na 12ª regra de “Acentuação Gráfica” do Formulário Ortográfico de 1943. Traz essa regra duas observações: uma, também em vigor ainda, que diz: “Não se põe acento agudo na sílaba tônica das formas verbais terminadas em *qüe*, *qüem*: *aproprinqüe*, *delinqüem*...”; outra, que perdeu vigência durante a sobremesa de um repasto acadêmico no dia 18 de dezembro de 1971; convivas lusitanos e brasileiros concordaram em aliviar-nos da licitude de tremar uma vogal “para indicar que seu encontro com outra não forma ditongo, mas hiato: *saudade*, *vaidade*”. Essas duas palavras – dadas oficialmente como exemplos no FO de 1943 – já não se apresentavam enfeitadas no Brasil; foi uma concessão dada de barato pelos portugueses para que se pudesse falar em realização de um “acordo”, e não de um simples banquete em nossa nova capital (ALMEIDA, 2001, p.569).

Essa definição torna-se inaplicável ao português brasileiro atual e contraria o princípio da inseparabilidade do ditongo, pois o trema que restou na língua indica que existe ditongo, justamente por a vogal “u” ser pronunciada como a semivogal /w/, de forma que não se trata de dígrafo. Como se percebe, a queda do trema foi gradativa na língua portuguesa e será total a partir da entrada em vigor do acordo ortográfico já aprovado.

O sistema de acentuação gráfica do português é eficaz, pois se orienta pela intuição do falante e tenta deixar o texto mais claro ao leitor. Dois exemplos disso são o trema e o “ê” seguido de “m” em sílaba travada no final de alguns vocábulos, pois o primeiro garante que o “u” sobre o qual está é pronunciado e o segundo é uma

sinalização morfossintática exclusiva de determinados verbos – "ter", "vir" e derivados – na terceira pessoa do plural (SCLIAR-CABRAL, 2003), indica, portanto, que se trata de plural.

O sistema de acentuação gráfica do Brasil baseia-se no princípio da economia de acentos, então, considera a intuição do usuário da língua. Assim, facilita a descodificação, embora pareça complicar a codificação, sobretudo, por não ser propriamente compreendido pelos professores e, conseqüentemente, ensinado aos alunos.

1.5.2.3 Clítico

Clíticos são, em sua maioria, palavras monossilábicas pronunciadas com pouca intensidade, ou seja, não têm acentuação própria. Os monossílabos tônicos soam distintamente no interior da frase, já os monossílabos átonos, os clíticos, soam como uma sílaba da palavra que os antecede ou precede. Então, fonologicamente, eles não são capazes de sustentar acento sozinhos, e precisam se agregar a um vizinho.

Em português, o clítico "me" apresenta essas características. Ele é sintaticamente ativo, já que é o objeto do verbo, fonologicamente, porém, ele é deficiente e precisa se agregar ao verbo, tornando o complexo verbo+clítico uma só unidade em termos fonológicos (GOMES, 2008).

Há os clíticos especiais e os locais. Os primeiros podem se agregar a qualquer palavra, desde que ela esteja na posição adequada. Os segundos agregam-se às palavras que os regem. Assim, alguns pronomes pessoais oblíquos, que são objetos de verbo ("me", "se", "lhe" etc.), são clíticos e, dada sua posição em relação ao verbo, podem ser proclíticos ("*se* mexer"), mesoclíticos ("dever-*se*-ia") e enclíticos ("fazê-*lo*").

A colocação dos pronomes oblíquos em relação ao verbo na história do português é fato curioso. Num primeiro momento (português arcaico, século XIII a XVI), ela foi predominantemente enclítica. Passou a ser proclítica no português clássico (século XVI ao XVIII) e voltou a ser enclítica no português contemporâneo (GALVES, 1996 *apud* NAMIUTI, 2008). Em nenhuma outra língua românica isso

aconteceu, pois todas eram enclíticas, passaram a proclíticas, e assim permaneceram (MARTINS, 1994 *apud* NAMIUTI, 2008).

Casos de clíticos são poucos em português e restringem-se principalmente a preposições, pronomes oblíquos, artigos e conjunções. Os pronomes oblíquos "comigo", "contigo", "conosco" e "convosco" resultam da combinação da preposição "com" com os pronomes oblíquos correspondentes, então, não são clíticos.

O acento grave em português indica a crase da preposição "a" com a forma feminina do artigo definido ("a", "as") e com os pronomes demonstrativos. Assim, o acento grave sobre a vogal "a" assinala a vogal átona /a/, pois se trata de um clítico, além do processo fonético da crase de /a/ mais /a/ e dos processos morfossintáticos de regência da preposição "a" sobre artigo, pronome substantivo ou pronome demonstrativo iniciado pelo fonema /a/.

Com exceção dos pronomes oblíquos, cuja posição pode ser proclítica, mesoclítica ou enclítica, a posição dos clíticos no português do Brasil é sempre proclítica (CÂMARA JR., 1997).

1.5.3 Variação Dialetal

É impossível abranger todas as variações dialetais e sociais da língua, pois ela é variável no espaço, na hierarquia social e num mesmo indivíduo. Com base nisso, um sistema descritivo deve comportar-se fonologicamente numa variedade que abranja as maiores possibilidades de realização lingüística (CÂMARA JR., 1986).

Para que a língua permaneça caracterizada, essas variações devem ser superficiais. A invariabilidade profunda, em meio a variabilidades superficiais, é inegável nas línguas, pois o princípio das invariantes nas variações é a chave da descrição lingüística, e ele cria o conceito de padrão, cuja depreensão numa língua dada é o objetivo central da gramática descritiva de tal língua (CÂMARA JR., 1986). Desse modo, as irregularidades, que há em toda língua, obedecem, em profundidade, a padrões particulares, que se coordenam com o padrão ou regra geral, dito regularidade.

Falantes de uma mesma língua diferem uns dos outros, mas há um padrão nuclear de comportamento (CÂMARA JR., 1986). Em um sistema eletrônico algumas dessas variações serão mais fáceis de inserir do que outras.

As variações de pronúncia na sociedade são freqüentemente objeto de julgamento de valor por parte dos falantes. É natural que em toda sociedade se estabeleça uma norma de pronúncia aceita e usada pela classe dirigente e, bem ou mal, imitada pela coletividade. Todo desvio de norma representa um valor que pode, por exemplo, simbolizar uma posição desprezada no grupo social. Os falantes que quiserem se livrar dessa marca de inferioridade têm de modificar sua pronúncia. À medida que essa organização toma a forma de ensino organizado (na escola) configura um exemplo de aplicação pedagógica da fonética, da mesma maneira que a descrição de uma norma prescrita (por escola, por academia) exige uma consideração consciente dos fatos fonéticos (MALMBERG, 1993).

Os dicionários normativos dão a pronúncia "correta" (padrão) em transcrição fonética. Todo ensino de língua materna na escola supõe por parte do professor conhecimento dos fatos fonéticos próprios dessa língua. A fonética, principalmente seu aspecto fonológico, é fundamental para o estabelecimento e a modificação (modernização) dos sistemas de escrita (alfabetos fonológicos) (MALMBERG, 1993).

Não se fala da mesma forma em todas as ocasiões. Quando o lingüista sincrônico se volta contra o gramático normativo, é em geral porque ambos confundem seus lugares na sociedade, e impõem suas regras praxistas como se fossem lingüística. A gramática normativa tem seu lugar, e não se anula diante da gramática descritiva, mesmo que esse lugar seja imposto por exigências de ordem prática na sociedade. Trata-se de duas disciplinas correlatas e independentes (CÂMARA JR., 1986).

A escolha da variação lingüística no ensino escolar é de certo modo predeterminada. A descrição não pode pautar-se em uma modalidade regional ou remotamente regional, nem se assentar em um uso elaborado e sofisticado, como a literatura. Ela deve partir do uso falado e escrito considerado culto, adequado às condições formais de intercâmbio lingüístico. É o que se pode chamar de lingüística aplicada a um comportamento social (CÂMARA JR., 1986).

Uma das finalidades do programa Nhenhém é contribuir no ensino escolar que, também por conta dessa rivalidade entre gramáticos e lingüistas, está prejudicado e não consegue combater o analfabetismo funcional.

1.6 MORFOLOGIA

A divisão do enunciado em morfemas (signos mínimos ou unidade lingüística que tem significante e significado) chama-se primeira articulação da língua (MALMBERG, 1993). Na primeira articulação da língua, em que o segmento fônico se associa a uma significação léxica ou gramatical, o vocábulo formal é a contraparte do vocábulo fonológico. Ao contrário do critério fonológico que rege a escrita, o qual procura representar aproximadamente os fonemas pelas letras e divide suas seqüências, a apresentação do vocábulo na escrita faz-se pelo critério formal. As unidades formais de uma língua podem ser livres e presas. São livres (independentes) quando constituem uma seqüência que funciona isoladamente como comunicação suficiente (palavras, seqüências coerentes de palavras), são presas (dependentes) quando funcionam ligadas a outras (os morfemas, como "pro" de "prover"). Tem-se o vocábulo formal quando não é mais possível dividir uma forma em duas ou mais formas livres – é uma forma livre indivisível ("bem"), ou duas ou mais formas presas ("bomba"), ou uma forma livre e uma ou mais formas presas ("desrespeito") (CÂMARA JR., 1986).

Cabem aqui considerações sobre alguns prefixos e radicais, como "troux", "trans", "sub" e "ob", cujo uso em vocábulos foge do padrão da língua. Normalmente, ao sucederem a letra "s" ocorrem os desvios. Os casos como "**ob**servar" e "**ob**séquio", "**sub**sídio" e "**sub**seção" já foram mencionados em outras seções.

O prefixo "trans" pode ser descodificado como /trãz/ ou /trãs/. Esse caso também já se abordou. Cabe complementar que, na escrita, há as duas possibilidades, como em "transa" → /trˈãza/ e "trança" → /trˈãsa/, em que são radicais. Então, não soam estranhas as duas opções ao falante. Dessa forma, não há regra sólida fonêmica para sua ocorrência.

Pela lógica, a leitura do radical “troux” deveria ser /trˈowʃ/, e não /trˈows/, uma vez que sucede ditongo, como em “frouxo”→ /frˈowʃu/. O uso consagrou essa descodificação anômala. Há outros casos assim em português.

1.7 SINAIS DE PONTUAÇÃO E SÍMBOLOS

Fazem parte da escrita da língua portuguesa vários sinais legíveis e não legíveis, como os sinais de pontuação de texto e símbolos.

Os sinais de pontuação são ponto final, ponto de exclamação, ponto de interrogação, vírgula, ponto e vírgula, dois pontos, reticências e travessão (.!?,;:...–). Num nível mais interno ao texto estão o hífen, que é usado para unir palavras, as aspas e o apóstrofo (-“”’). Muitos dos sinais de pontuação usados em português já eram usados em latim.

O hífen tem várias utilidades na escrita. Ele é usado como traço de união em algumas palavras compostas, como sinal de subtração, como travessão, como indicador de itens de lista, e pode haver mais usos.

Quanto aos símbolos, há parêntese, chave, colchete, cifrão, arroba, percentual, “e” comercial, asterisco, sinal de soma e de divisão, sinal de parágrafo, sinal de igual, de maior e de menor, abreviatura de número, de feminino, barra, barra invertida, barra vertical, *underscore*, cerquilha (){}[]&@%&*+÷§=><°ª/_#|). Ainda há outros menos usados. Os números indo-arábicos também são símbolos.

1.8 COMPUTAÇÃO

A partir da eletroeletrônica, o homem desenvolveu o computador e, juntamente com ele, uma linguagem própria de máquina, para que houvesse interação entre ambos. Um computador digital é uma maquina capaz de solucionar problemas por meio da execução de instruções que são fornecidas a ela. Esse conjunto de instruções é chamado de programa. Para o computador decifrar essas instruções, elas têm de ser simplificadas, de forma que possam ser reconhecidas pelos circuitos lógicos. É o

processamento eletrônico. Assim, os programas têm de ser convertidos para uma linguagem de máquina. Por meio dos programas, se estabelece a comunicação final entre homem e máquina.

1.8.1 Descodificação Eletrônica

O computador trabalha com números, ou seja, internamente, ele relaciona qualquer caractere (letra, expressão numérica, outros sinais gráficos) e comando (salvar e abrir arquivo, trocar de linha em um texto) a números específicos. Para isso, ele se baseia em um sistema codificador, que atribui um código único a cada caractere, independentemente da plataforma operacional em que esses caracteres estejam. Fundamentalmente, o computador trabalha com sinais elétricos. Todo comando é convertido em pulsos elétricos. Assim, pulsos fortes são considerados 1 e pulsos fracos ou ausência de pulsos, 0. Isso associado a um intervalo de tempo fornece seqüências de oito em oito *bits*,[23] que correspondem a um *byte*,[24] que corresponde a um caractere ou a um comando.

Os sistemas numéricos usados nessa codificação são binário (dois dígitos = 0 e 1), o único em que o computador trabalha, decimal (dez dígitos = 0 a 9) e hexadecimal (16 dígitos = 0 a 9, A, B, C, D, E, F), usados pelo homem para criar códigos exclusivos para comunicar-se com a máquina.

A tabela ASCII[25] foi o primeiro codificador de caracteres para computador, e sua base ainda é mantida nos computadores. A partir dela, outros codificadores foram criados, sobretudo para representar eletronicamente caracteres lingüísticos. Tal representação exigia mais códigos do que o sistema ASCII continha. Os computadores de hoje usam sistema hexadecimal ASCII ou Unicode.

---

[23] Abreviatura de *binary digit* (dígito binário), que pode ser 1 ou 0: sim ou não, verdadeiro ou falso.
[24] Abreviatura de *binary term* (termo binário), que corresponde a uma seqüência de oito *bits*.
[25] Sigla de American Standard Code for Information Interchange.

1.8.2 Código ASCII

A tabela ASCII surgiu da necessidade de fazer as máquinas interagirem, ou seja, ler dados e descodificá-los da mesma maneira. Afinal, para criar uma rede de computadores são necessárias máquinas compatíveis e, mais importante, um sistema de códigos gráficos (uma espécie de alfabeto) comum, que todas as máquinas possam decifrar igualmente.

Antes de maio de 1961, a maioria dos sistemas de computadores tinha uma maneira particular, individual, de representar caracteres alfanuméricos. Foi proposto o uso de um código comum, a fim de possibilitar a comunicação entre os computadores e permitir o intercâmbio de dados entre máquinas de diferentes tipos e fabricantes. Como um alfabeto desse tipo não existia, formou-se um comitê de representantes da indústria e do governo dos Estados Unidos para estudar o assunto. O resultado do trabalho desse comitê foi o primeiro padrão universal para computadores, que foi chamado de Código Padrão Americano para Troca de Informações, o código ASCII.

ASCII é um código numérico usado para representar os caracteres, entendido por quase todos os computadores, impressoras e programas existentes. Ele é baseado no alfabeto romano, como é usado no inglês moderno, e visa a padronizar a forma como os computadores representam letras, números, acentos e sinais diversos (por exemplo: <, {,]) e alguns códigos de controle (como <Crtl> e <Alt>), que são utilizados para converter todos os símbolos em números binários, os quais efetivamente podem ser processados. Na tabela ASCII, letras minúsculas, maiúsculas e com diacríticos têm valores (códigos) diferentes.

Tabela 1: Exemplos de código no sistema binário e decimal

		Sistema decimal (base 10)
Caractere	**Sistema binário**	**Código ASCII**
a	01100001	97
A	01000001	65
ã	11100011	227

Na primeira tabela ASCII havia 95 caracteres imprimíveis. Eles eram numerados de 32 a 126, pois os primeiros códigos (de 0 a 31) foram reservados para caracteres de controle, ou seja, que controlam funções ou equipamentos. Esses caracteres de controle originaram-se nos primórdios da computação, quando eram utilizadas máquinas Teletype (como máquinas de escrever eletromecânicas), fitas de papel perfurado e impressoras de cilindro, portanto, muitos deles eram dirigidos a esses equipamentos (UFPA, 2007).

A primeira tabela ASCII era um conjunto de 128 números únicos (de 0 a 127), expressos em 7 *bits*, que correspondiam a cada letra do alfabeto latino, a cada um dos algarismos arábicos, a diversos sinais de pontuação e a algumas funções especiais (como quebra de linha e retorno de carro). Até hoje, qualquer computador contém essa tabela, de modo que, quando se digita no teclado da máquina, ela descodifica as teclas digitadas, e mostra as letras certas no monitor.

Essa tabela foi criada com 128 números por causa do padrão de 8 *bits* para um *byte* da International Business Machines Corp. (IBM), que essa empresa conseguiu impor em 1964, e que tornou obsoletos os padrões de outras empresas. A relação entre 8 *bits* e 128 números, é que, no sistema binário, usando-se todos os 8 *bits*, é possível obter 256 valores diferentes ($2^8 = 256$). Como o oitavo *bit* havia sido reservado como *bit* de paridade, importante nas comunicações, restavam 7 *bits* para guardar números, e $2^7 = 128$. O *bit* de paridade era um dígito verificador de controle do código digitado, a exemplo dos dígitos do CPF. Com o tempo, os 128 códigos se mostraram insuficientes e abandonou-se o *bit* de paridade, o que aumentou o número de códigos para 256. É a chamada tabela ASCII estendida ou expandida.

Quando os primeiros computadores foram projetados, percebeu-se que seriam necessários cerca de 250 códigos diferentes para representar, com valores diferentes, todos os números, letras maiúsculas, minúsculas e acentuadas e os demais símbolos. Assim, no sistema de caracteres ASCII, cada valor binário entre 0 e 127 está associado

a um caractere específico. Os últimos 128 códigos comportam elementos especiais, como caracteres acentuados em diferentes línguas, como o português.[26]

Os valores ASCII são usados no sistema decimal, hexadecimal e binário. Cada caractere de todas as fontes de texto para computador tem correspondência em ASCII, inclusive as fontes usadas em fonética. Em ASCII decimal, os códigos vão até 255, em hexadecimal, até FF. O código ASCII é o código mais usado na comunicação entre computadores de diferentes tipos e fabricantes.

1.8.3 Código Hexadecimal

Como se mencionou, o sistema hexadecimal utiliza códigos de 0 a 9 e letras de A a F, para compor eletronicamente os símbolos gráficos e comandos de que o homem precisa para comunicar-se com o computador. O Unicode é um sistema hexadecimal codificador eletrônico de caracteres que abrange sozinho os caracteres que representam as inúmeras e diferentes línguas faladas no mundo. Dessa forma, não há conflito de código, ou seja, o mesmo código não é atribuído a mais de um caractere nas mais diferentes línguas, em qualquer máquina, sistema operacional e plataforma (UNICODE, 2007). No entanto, sabe-se que, em computação, sempre pode haver certo grau de incompatibilidade entre computadores, linguagens de programação e programas, e isso gera conflito de código.

Além disso,

> The Unicode abstract character repertoire can, in theory, hold up to 1114112 characters, although many are reserved to be invalid and the rest aren't all likely to ever be assigned. Each character is coded as an integer between 0 and 1114111 (0x10ffff). […] The `char` is the most basic character type. Each `char` is a single Unicode character. It takes 2 bytes in memory, and can take a value of 0-65535. Note that not all values are thus actually valid Unicode characters (YODA, 2007).

[26] Embora o sistema de acentuação do português do Brasil facilite a descodificação pelo usuário, na computação, o inverso é verdadeiro. Para o computador, os acentos são mais problemáticos para ser codificados.

Se cada caractere no sistema hexadecimal Unicode ocupa até dois *bytes* e cada *byte* ocupa 8 *bits* de memória, então, cada caractere pode ocupar até 16 *bits*. No entanto, usam-se apenas 8 *bits* no sistema hexadecimal, para dar conta dessa mesma situação. Esse sistema divide os oito *bits* em dois, de forma que cada grupo de quatro bits corresponde, respectivamente, à letra e ao número que formam o código. Os códigos 97, 65 e 227 correspondem às letras "a", "A" e "ã" em ASCII decimal, como visto, e para essas mesmas letras, nos sistemas hexadecimais, tem-se:

Tabela 2: Exemplos de código no sistema hexadecimal

		Sistema hexadecimal (base 16)
Caractere	**Sistema binário**	**Código ASCII**
a	01100001	61
A	01000001	41
ã	11100011	E3

1.8.4 Fontes do IPA e SIL

O International Phonetic Alphabet (IPA) é a maior e mais antiga organização representativa para foneticistas, criada em 1886, em Paris. O IPA fornece padrões para representações fonéticas para todas as línguas (IPA, 2007). O Summer Institute of Linguistics (SIL), uma organização cristã, cujo foco é estudar línguas que não são escritas, desenvolve programas de computador para estudá-las (SIL, 2007).

As fontes do IPA (Anexo 1) e do SIL obedecem ao código hexadecimal, caso contrário, não poderiam ser descodificadas em qualquer computador. Por conseqüência, os caracteres escritos do alfabeto latino moderno, nas fontes do IPA, usam os mesmos números da tabela ASCII e adaptam caracteres lingüísticos especiais.

As fontes do IPA foram feitas com base na língua inglesa (IPA, 2007), então, pressupôs-se que as transcrições seriam feitas nessa língua, o que dificulta seu uso em fonética e fonologia de outras línguas, como o português, que usa sinais na transcrição que não fazem parte das transcrições da língua inglesa. Nesse caso, decisões devem ser tomadas em favor de uma transcrição legível e possível, com os recursos disponíveis.

Na fonte SILDoulos IPA 93, por exemplo, faltam caracteres para representar o arquifonema |S| do português, porque seu código em hexadecimal (0053) corresponde a /ʃ/, que representa outro fonema. Da mesma forma ocorre com o arquifonema |W|, cujo código ASCII hexadecimal (0077) equivale a /ʷ/. O /w/ já é usado para representar semivogal em português. Outros problemas de correspondência referem-se ao IPA usar um código para um caractere especial totalmente diferente do código em ASCII, o que impede que se use o caractere original, como o caractere 00E3, que em ASCII (227) corresponde a "ã" e em IPA, a /ɯ/. Nas fontes do IPA, letras com diacríticos são representadas por sinais compostos. Para se obter /ã/, no IPA, precisa-se digitar a+~ (0061+0029). Isso pode gerar erros de compatibilidade entre programas eletrônicos de um mesmo computador e, é claro, em computadores diferentes, mesmo que as fontes requeridas estejam instaladas.

As fontes do IPA diferem entre si, em alguns casos, em que os mesmos caracteres têm códigos distintos, e também algumas fontes contêm mais caracteres do que outras. A fonte IPAPhon parece ser mais completa, pois traz, por exemplo, caracteres minúsculos e maiúsculos das letras, que, então, apresentam códigos diferentes. Há os caracteres /S/, /s/, /ʃ/ e /r/, /ʀ/, /ʁ/, que correspondem respectivamente aos códigos 53, 73, A7 e 72, 52, 91. Isso facilita o uso dessa fonte, porque ela oferece mais opções. Apesar disso, testes com o caractere 91 mostraram que ele pode ser ilegível por algumas máquinas e linguagens de programação. Versões diferentes de editores de textos também podem mostrar incompatibilidades nos caracteres especiais.

O IPA é o alfabeto fonético de maior uso por pesquisadores treinados em registrar as diferentes falas (SCLIAR-CABRAL, 2003).

1.8.5 Lógica de Programação

Na informática, lógica de programação é a forma pela qual assertivas, pressupostos e instruções são organizados em um algoritmo para implementação de um programa de computador. Algoritmo é um conjunto de regras e operações bem

definidas e ordenadas, destinadas à solução de um problema, em um número finito de etapas.

Fundamentalmente, um programa de computador é um conjunto de instruções que permitem resolver um problema. Criar um programa desses consiste em desenvolver e agrupar comandos que permitam ler uma informação (entrada), processá-la eletronicamente e fornecer uma resposta desejada (saída). Para isso, é preciso informar à máquina as ações a executar (instruções), o que é feito com a digitação de estruturas lógicas. Uma das estruturas elementares da programação de computadores é a condicional

Se... então... senão...
ou
If... then... else...

Veja-se, como exemplo, o algoritmo de um programa simples.

```
SE A > B ENTÃO
ESCREVA('A é maior do que B')
        SENÃO
        SE A < B ENTÃO
        ESCREVA('A é menor do que B')
                SENÃO
                ESCREVA('A é igual a B')
```

Se o último **senão** não fosse feito, o programa geraria um erro, porque poderia ocorrer uma situação imprevista.

Para trabalhar com essa estrutura lógica, o computador precisa comparar valores, que são armazenados como variáveis. O computador tem uma área de armazenamento conhecida como memória. As informações existentes no computador estão na memória primária ou principal (memória de acesso aleatório – RAM) ou na memória secundária (discos, CD-ROM etc.). A memória do computador pode ser entendida como uma seqüência finita de caixas que, num dado momento, guardam algum tipo de informação, como número, letra, palavra, frase etc. Importa que lá

sempre existe alguma informação. Nesse sentido, uma variável é uma posição de memória, representada por um nome simbólico atribuído pelo programador, a qual contém, num dado instante, uma informação.

Para funcionar, a seqüência lógica anterior precisa de duas variáveis (A e B), que precisam ser de algum tipo (número inteiro, no caso):

```
PROGRAMA Maior
VARIÁVEIS
A, B: INTEIRO
        INÍCIO
                SE A > B ENTÃO
                ESCREVA('A é maior do que B')
                    SENÃO
                    SE A < B ENTÃO
                    ESCREVA('A é menor do que B')
                            SENÃO
                            ESCREVA('A é igual a B')
                                    SENÃO
                                    ESCREVA('Tipo incompatível')
        FIM
```

A condição acrescentada prevê o erro de o usuário digitar uma letra, por exemplo, uma vez que o programa espera um número.

Em resumo, um programa de computador tem nome, variáveis e rotinas (seqüências lógicas de procedimentos ou funções que têm início e fim). Um programa de computador funciona com lógica, o que subentende previsibilidade, ou seja, regras que prevêem acontecimentos e o que fazer quando eles ocorrerem. Regras habilitam o programa a lidar com todos os casos idênticos. Se não houver como criá-las, a solução é fazer bibliotecas ou listas de entradas que o programa não processa, apenas repete, são respostas prontas. Isso diminui sua eficiência.

Desse modo, se as exceções sempre foram o grande embaraço da regulamentação gramatical (CÂMARA JR., 1986), para um sistema eletrônico, elas constituem barreiras de igual proporção, pois o computador não trabalha com exceções, apenas com regras. Então, é viável que haja regras para as exceções.

2 ASPECTOS TÉCNICO-CIENTÍFICOS

cansei da frase polida
por anjos da cara pálida
palmeiras batendo palmas
ao passarem paradas
agora eu quero a pedrada
chuva de pedras palavras
distribuindo pauladas

Leminski (2008)

A metodologia que o programa desenvolvido como objeto desta pesquisa segue em suas regras intrínsecas vem da adaptação da teoria exposta na seção anterior, uma vez que ela serviu de base para que se tomassem decisões fundamentais para o funcionamento do Nhenhém. Quando a teoria transforma-se em prática, surgem incompatibilidades que exigem resolução, porque o programa criado tem de verificar tudo o que o usuário quiser, dentro de sua finalidade, que é descodificar grafofonemicamente a língua portuguesa escrita do Brasil.

A necessidade de adotar padrões para evitar inconsistências e erros faz recriar teorias, tomar partido e documentar ações. Desse modo, a metodologia, sobretudo no tocante a massa lingüística fonológica, faz parte do conteúdo da tese.

2.1 LINGÜÍSTICA COMPUTACIONAL

A lingüística computacional pode ser entendida, de modo restrito, como o desenvolvimento e a aplicação de recursos eletrônicos para manusear propriamente dados lingüísticos. Há alguns programas informatizados específicos para isso,[27] embora não sejam muitos os de fácil acesso aos pesquisadores, porque o idioma em

[27] Um exemplo de programa eletrônico para ler dados lingüísticos é o WordSmith©, feito em língua inglesa, cujo funcionamento foi descrito por Gerber (2007). Esse programa lê os textos em sua forma escrita e fornece estatística sobre ocorrências de vocábulos, mas não processa nem traduz fonologia e fonética. Outro exemplo é o analisador sintático eletrônico da língua portuguesa Palavras (SANTOS, BICK, MARCHI e AFONSO, 2007), que está disponível na Rede Mundial.

que trabalham é estrangeiro, por não estarem acessíveis nos ambientes de pesquisa e por sua correta utilização exigir treinamento nem sempre disponível. Sistemas computacionais específicos para trabalho com fonética e fonologia são mais raros ainda.

O fato de um sistema descritivo dever comportar-se fonologicamente numa variedade que abranja as maiores possibilidades de realização lingüística alia-se à computação e facilita o desenvolvimento de um sistema descritivo eletrônico da língua, no que tange à conversão fonológica, embora seja possível escolher uma variedade padrão e digitalizá-la completamente. Isso depende de o que se deseja extrair dele.

O desenvolvimento de modelos computacionais da língua permite maior processamento de informações, tendo em vista que a maior parte do conhecimento está registrada na forma lingüística (OLIVEIRA, 2003).

2.1.1 Processamento Eletrônico da Língua

Um sistema de processamento da língua funciona com entradas e saídas, e a linguagem de programação utilizada para o processamento computacional é responsável pela interface requerida. Assim, o sistema Nhenhém funciona da seguinte forma, como qualquer sistema eletrônico:

Entrada	→	**Processamento**	→	**Saída**
Gráfica e em língua portuguesa [escrita]		Em linguagem de programação de computador		Gráfica e em língua portuguesa [representação da fala]
Sílaba Palavra Frase Sentença Texto	→	Conversão de caracteres em símbolos do IPA e deles em números	→	Relatórios de conversão e de características fonológicas, levantamentos numéricos

Figura 3: Esquema de processamento computacional

Desse modo, para um sistema computacional tratar a língua, devem-se satisfazer duas condições: a entrada e a saída do sistema devem ser codificadas em uma língua e

o processamento da entrada e a geração da saída devem ser baseados em aspectos estruturais dessa língua.

Um sistema computacional trabalha com rotinas predefinidas, ou seja, o computador somente decifra aquilo para o que é preparado, de modo que é preciso instruir cada passo que o programa dará e prever e resolver os erros que ocorrerão. Qualquer ocorrência não prevista gerará um erro.[28]

Assim, quanto mais delimitações houver no assunto que o sistema tratará, menos custoso será seu desenvolvimento e mais garantido será seu desempenho. Em se tratando de processar a língua, a falta de delimitações é uma imensa barreira. No entanto:

> Os lexemas de uma língua existem em número muito elevado [...]; os elementos gramaticais existem em número limitado e os fonemas, finalmente, representam um estoque de unidades muito restrito (cerca de cinqüenta). É evidente que o número de enunciados (textos) é ilimitado. O princípio da linguagem humana está, portanto, na possibilidade de construir, com base em um número restrito de elementos mínimos, um número ilimitado de realizações lingüísticas (enunciados, textos) (MALMBERG, 1993, p.9).

No que tange à transcrição, é mais fácil desenvolver um sistema eletrônico fonológico do que um sistema fonético, dada a complexidade de símbolos que a fonética envolve. No entanto, quando se trata de extrair informações numéricas dos dados, a fonologia torna-se menos flexível, pois há necessidade de haver fronteiras mais claras do que as que ela apresenta.

---

[28] Isso não significa que o computador é "burro", essa metáfora é errônea, válida para quem tem capacidade de aprender, mas não consegue. O computador é lógico, o ser humano é racional. Esperar racionalidade de uma máquina lógica não é ser racional. Percebê-la como "burra" parece desculpa para quem não sabe usá-la. Homens pensam (isso é complexo e pode ser imprevisível), máquinas apenas processam (isso é simples e previsível). Máquinas são instruídas e somente respondem de acordo com instrução prévia. Entender isso é básico para trabalhar bem com elas e delas extrair respostas satisfatórias. Diante de um erro, um programador de computadores não diz que a máquina é burra, ele se pergunta onde falhou na instrução.

2.2 LINGUAGEM DE PROGRAMAÇÃO DELPHI

A linguagem de programação em que o aplicativo foi feito não é específica para pesquisas da língua, pois o Nhenhém não se diferencia de outros sistemas de gerenciamento de informações.

Uma das linguagens de programação mais utilizadas desde os anos 1990 é Delphi©, da Borland, que pode ser considerada o aprimoramento do Turbo Pascal, o qual foi vastamente usado nos anos 1980 e início dos anos 1990. Trata-se de uma linguagem de programação de computadores orientada a objeto, que roda em ambiente Windows©. Então, essa linguagem permite contato com o ambiente virtual mais utilizado pelos usuários de computador e é munida de ferramentas e estilos conhecidos deles, o que torna o sistema autoexplicativo e contribui para a interação usuário-sistema. O Delphi está na versão 10, mas sua versão mais popular ainda é a 7.0, na qual o Nhenhém foi desenvolvido.

2.3 FONTE IPAPhon

No sistema Nhenhém, tudo deve ser convertido para caracteres do IPA. Nesse caso, o uso de arquifonemas ficou limitado na fonte SilDOulos IPA 93, uma vez que ela provoca conflito com grafemas usados na escrita do português. Por exemplo, o arquifonema |S| provoca conflito com o fonema /ʃ/, porque nas tabelas internas de caracteres usadas pelo computador (sistema decimal e hexadecimal – tabela ASCII), ambas têm o mesmo código (53), apesar de se tratar de fontes (tipo de letra) diferentes, pois |S| pode aparecer em qualquer fonte *true type* e /ʃ/ faz parte das fontes do IPA.

Nesse sentido, a fonte IPAPhon atende mais bem às necessidades, pois contém três tipos de “s”, que funcionam distintamente. Ela também contém os tipos de “R” que seriam necessários para suprir necessidades do sistema, mas um deles mostra-se incompatível com computadores pessoais: o já mencionado caractere 91. Nesses casos, a melhor solução é usar outros caracteres ou combinações de caracteres que sejam legíveis.

2.4 LINGÜÍSTICA DE *CORPUS*

A lingüística de *corpus* é a forma apropriada para lidar com massa lingüística. Assim, sempre que se trabalha com *corpus*, devem-se levar em conta as prerrogativas dessa teoria. Ela envolve a lingüística computacional e a estatística, uma vez que sempre se trabalhará com o computador e com uma amostra da língua. Respeitando-se o que aconselha a teoria sobre trabalho com *corpus*, documentada em Sinclair (1991), Leech (1996), McEnery e Wilson (1997), Rocha (2000), Sardinha (2000), considera-se que qualquer texto que se encaixe em seus pressupostos pode ser fonte de investigação.

A lingüística de *corpus* remonta e recupera práticas consagradas há séculos, que hoje voltam com toda a carga, para desafiar ortodoxias hegemônicas que se encontram enraizadas no campo da lingüística, a partir de sua consolidação como disciplina independente. Dados concretos são convocados apenas para confirmar um conceito universal da língua, para servir de exemplificação de modelos ou hipóteses preestabelecidos (RAJAGOPALAN, 2007).

A lingüística de *corpus*, como uma das mais promissoras teorias lingüísticas da atualidade, não deve desconsiderar qualquer área que trate da língua, nem mesmo a gramática normativa. Ela deve, sim, desfrutar o que delas é pertinente para a pesquisa socioacadêmica e renunciar ao que não importa a seus propósitos, sem preconceitos de qualquer ordem. Afinal, o valor de uma teoria comprova-se em sua prática. A experiência de construir o Nhenhém mostra que essa realidade pode ser mais intensa no campo da lingüística computacional.

2.4.1 Estatística

Quando se fala em massa de dados, subentende-se que se utilizará estatística, tendo em vista que é inviável verificar individualmente milhares de combinações de fonemas. Autores da área são unânimes em corroborar a afinidade entre a lingüística de *corpus* e a estatística (MANNING e SCHÜTZE, 2000). Por definição, estatística é o ramo da matemática que trata da coleta, análise, interpretação e apresentação de

grandes quantidades de dados numéricos, geralmente com o propósito de inferir as proporções em um todo, a partir das proporções de uma amostra representativa. Assim, estimam-se parâmetros (características descritivas dos elementos da população),[29] por meio de estatísticas (operações com dados da amostra) (BARBETTA, 2000). A probabilística é a parte da estatística que lida com a perspectiva favorável de fatos acontecerem.

Uma das características da abordagem com base em *corpus* é que ela depende de técnicas de análise quantitativas e qualitativas, uma vez que interpretações qualitativas dos dados quantitativos são essenciais para explicar as razões por que há certos padrões de ocorrência, por exemplo. Análises quantitativas estão intimamente ligadas a estudos com base em *corpus*, pois, havendo bastantes dados, é possível obter freqüências e aplicar testes estatísticos, bem como dispor os resultados em gráficos e fazer várias comparações (BIBER, CONRAD e REPPEN, 1998).

A estatística fonética foi aplicada à estenografia desde o início do século XX. As grafias mais simplificadas se aplicam aos elementos mais freqüentes, aqueles cuja previsibilidade é maior. Da mesma maneira, a distribuição das letras do teclado do computador – que é herdada da máquina de escrever –, para ser racional, demanda conhecimento análogo a esse. No teclado, as letras devem estar dispostas de maneira tal que as mais freqüentes exijam o mínimo de movimento (MALMBERG, 1993).

A estrutura da expressão é simples comparada à do conteúdo, que compreende na verdade toda a experiência humana material e espiritual de hoje, ontem e de amanhã. Os morfemas praticamente existem em número ilimitado, e esse número aumenta incessantemente com a introdução de novas palavras e novas maneiras de se expressar, criadas no interior da língua ou provenientes de outras línguas, por empréstimo ou imitação. A estrutura fonológica é muito mais estável, muito mais restrita. Os empréstimos são raros e de pouca importância (MALMBERG, 1993).

A descrição da expressão, à medida que se concentra nos valores funcionais (fonologia) ou nas características sonoras ou físicas (fonética pura) é qualitativa, mas

[29] População é o conjunto de elementos abordados pelo estudo – ou seja, qualquer entrada do sistema – para os quais se deseja que as conclusões obtidas sejam válidas; enquanto amostra é a seleção de parte da população para ser observada (BARBETTA, 1998).

os elementos da expressão têm também o aspecto da quantidade ou freqüência de ocorrência. Trata-se de um elemento freqüente ou raro num enunciado, e há regras fonéticas para essa freqüência. Por exemplo, em uma língua que contém consoantes surdas e sonoras, as surdas são duas vezes mais freqüentes do que as sonoras. Da mesma forma, as sílabas abertas são muito mais freqüentes na língua do que as sílabas fechadas, que não existem em várias línguas. As palavras breves são mais freqüentes do que as longas em qualquer enunciado. Qualquer desvio dessas regras gerais de distribuição e ocorrência implica acréscimo de um fator à informação. São os fatos de estilo, que enriquecem a mensagem (MALMBERG, 1993).

Não se trata de mero acaso, pois fatos assim revelam um princípio extremamente importante no mecanismo da expressão e da língua. Há uma relação positiva entre simplicidade estrutural e freqüência de ocorrência: é o princípio da economia, ou seja, utilizar o máximo de elementos simples antes de recorrer aos elementos complexos (MALMBERG, 1993).

De maneira geral, toda ciência tradicional da língua tinha baseado seus resultados na coleta de fatos de acordo com o espírito do positivismo e supunha, portanto, apoio em procedimentos quantitativos e estatísticos. As descrições da fonética tradicional eram valores médios, assim como as medidas acústicas. Resultados cientificamente obtidos exigem, nas medidas, diferenças estatisticamente seguras, significantes. Na teoria da informação, cujo papel é considerável em certos ramos da lingüística e da fonética, os cálculos estatísticos e métodos matemáticos assumiram lugar central (MALMBERG, 1993).

O que os estudiosos denominaram fonometria implica uso mais conseqüente do método estatístico no campo da fonética. A aplicação da fonética em foniatria desde os anos 1920 pôs o problema do conceito da articulação normal e do valor normal da variação, em função dos quais os fatos anormais podiam definir-se. Nos anos 1920 e 1930, realizou-se uma série de pesquisas estatísticas de variação sobre os fenômenos fonéticos – respiração, intensidade, melodia, diferenças de duração etc. As mais conhecidas delas foram feitas pelos precursores alemães Eberhard e Zwirner (MALMBERG, 1993).

Como dito, em toda língua há regras específicas que limitam a utilização dos elementos fonéticos ou outros. Nem todas as línguas têm grupos consonantais, e nas que os contêm é normal, por exemplo, haver grupos iniciais como “tr”, “pr”, “kr”. A princípio, não existe língua em que haja “rt”, “rp” e “rk” em início de palavra, no entanto, em final de sílaba, essas combinações existem em inúmeras línguas. A maioria das línguas românicas é pobre em grupos finais de palavras e de sílabas. O número de oposições é, então, reduzido em relação à posição inicial. O estudo das regras de distribuição dos fonemas nas sílabas e nas palavras é uma fonotaxe, em analogia com a sintaxe, que é o estudo das combinações dos signos nos enunciados (textos) (MALMBERG, 1993).

O lingüista dinamarquês Hjelmslev apresenta um ponto de vista oposto ao dos estatísticos, por ser puramente qualitativo, no sentido de que a freqüência de ocorrência dos fatos não interessa, mas sim a existência deles, ou seja, a possibilidade de sua utilização. Então, se um fato paradigmático ou um tipo de sintagma for representado por um único exemplo, é preciso atribuir-lhe o mesmo lugar no sistema do que aquele ocupado por fenômenos de freqüência elevada (MALMBERG, 1993).

A estatística auxilia na sistematização de regras gerais da língua portuguesa e dos referidos desvios, bem como de seus efeitos. Por isso, após a conversão dos textos, o Nhenhém avalia-os em números inteiros e porcentuais. Quando não se usam casas decimais, 0% corresponde a valor inferior a 0,6% e 1% corresponde a valor entre 0,6% e 1,5%. É como se processa o arredondamento numérico no computador.

2.4.2 Massa Lingüística Fonológica

Embora os princípios da lingüística de *corpus* sejam válidos para qualquer *corpus* usado para pesquisar a língua, a lingüística de *corpus* não trata especificamente de fonologia. O usuário do aplicativo Nhenhém pode montar um *corpus* fonológico, isto é, de textos convertidos, dessa forma, critérios específicos para trabalho com *corpus* em fonologia devem ser pesquisados. A teoria geral sobre montagem de *corpus*

deve ser seguida e complementada. A exposição a seguir é fruto da experiência com o Nhenhém.

Um *corpus* para pesquisa fonológica contém a pronúncia de uma língua, ou seja, uma variedade de fala considerada geral, descodificada em grafofonemas. O que não é pronunciável não deve fazer parte dele, nem de qualquer texto isolado convertido. A partir disso, sinais de pontuação (.,;!?:’“” e outros) e caracteres especiais (/ºª@&$( )+[]{}-=* etc.) não fazem parte da conversão dos textos do *corpus*. Da mesma forma isso ocorre com algarismos indo-arábicos e romanos. Isso se aplica também a siglas, pois a sigla é pronunciada pelo nome das letras (CNPq: “Ceenepequê”) ou como uma palavra (BESC: “Besque”). A abreviatura “etc.” deve ser trocada pela palavra “etcétera”. Letras maiúsculas não têm efeito na pronúncia, mas o computador as distingue das minúsculas, o que implica códigos binários totalmente diferentes dos esperados e, conseqüentemente, tradução incorreta por erro de comparação. O programa não aceita digitação em letras maiúsculas e as lê como minúsculas em textos colados. No entanto, cabe atenção especial a esse caso na montagem do *corpus*. Sinais de pontuação e caracteres especiais são eliminados automaticamente, já as palavras e letras impróprias devem ser tratadas pelo usuário.

A conversão baseia-se em um alfabeto ortográfico e em um alfabeto específico fonológico, que concordam com o funcionamento do computador, do programa e das fontes do próprio alfabeto fonético (Anexo 2). Somente se analisa a entrada que tem correspondência nesse alfabeto específico de saída. Os caracteres que não se enquadram nessa condição devem ser ignorados pelo programa ao gerar estatística, para não haver erro nos cálculos.

O formato de arquivo ideal para os textos do banco de dados fonológico é .txt, a exemplo do *corpus* do NILC (2003), que pode ser lido a partir do Bloco de notas do Windows, em que o texto não tem formatação elaborada (cores, figuras, tabelas, marcadores e numeração, tabulação, paragrafação, capitulação, efeitos especiais de fonte). Com formatação simples, o texto fica leve, e se torna fácil carregar na memória do computador milhares de caracteres distribuídos em vários textos. Podem-se carregar

mais textos, o computador não trava com facilidade, mesmo que não seja de última geração.

O uso de fontes especiais, as do IPA, mostrou que é necessário que o editor de textos controle os caracteres da fonte, o que não pode ser feito pelo Bloco de notas, por exemplo, porque ele sempre abre com a última fonte usada pelo usuário do computador. O editor Word© mostrou-se mais eficiente para ler os textos convertidos, porque abre o texto com a fonte e o tamanho de fonte em que esse texto foi feito originalmente. No entanto, para montar um *corpus*, o formato seguro para a linguagem de programação Delphi é o.txt, pois a linguagem não consegue entrar com facilidade num texto .doc e fazer operações de checagem e reconversão dos caracteres, sobretudo se estão em fonte especial, como as do IPA.

Mesmo no formato .txt, há incompatibilidade entre as fontes básicas do computador e as do IPA, por causa das tabelas internas de caracteres. O computador não consegue ler universalmente um texto com fonte IPA. Por mais que pareça conseguir, pode haver incompatibilidades não perceptíveis à primeira vista. Estudos futuros específicos e aprofundados podem revelar formas de superar esse problema. No entanto, em lógica, há vários caminhos alternativos.

Uma solução é separar textos para gerar estatística e textos para ser manipulados pelo usuário. Assim, o usuário faz os ajustes de conversão e informa ao programa que aquele texto será usado para estatística. Depois disso, não se pode editar o texto. Se esse não for o caso, o usuário grava o texto em formato .doc, para ser aberto no editor de textos Word. O texto para estatística é manipulado internamente pelo programa, a fim de acertar as incompatibilidades. Isso garante a confiabilidade dos textos do banco e, conseqüentemente, a extração de resultados confiáveis. O programa grava os arquivos para análise em um formato próprio (.vve e .vvo).

Comportamentos lingüísticos raros mas canônicos não podem ser desconsiderados, no entanto, comportamentos raros e não canônicos ou não unanimemente aceitos pelas teorias podem ser excluídos da análise, porque a estatística necessita de muitas ocorrências do mesmo tipo, para ser válida. Esse seria o caso do /i/ epentético que pode ocorrer no encontro raro entre certas consoantes, como em

"confecção" → /kõfekisˈãw/, se fosse usado um símbolo que não é um fonema para sua representação, como [ɪ] ou [ʲ]. No entanto, como essa epêntese assemelha-se ao /i/ pretônico e postônico (CÂMARA JR., 1986), optou-se por usar a vogal fonêmica correspondente para representá-lo. Assim essa epêntese entra na estatística do fonema /i/.

A marca de tonicidade /ˈ/ também não precisa ser considerada por estatística que se baseie em contagem de fonemas, afinal, acento não é fonema. Essa marca pode ser útil em estatísticas mais elaboradas, em que se considerem posições nas palavras. Uma estatística assim informaria com exatidão, por exemplo, a freqüência de vocábulos oxítonos, paroxítonos e proparoxítonos na língua; de tonicidade de cada vogal; dos fonemas que margeiam a vogal tônica. Esse tipo de análise não era previsto no projeto desta tese, bem como estatística sobre encontros consonantais e vocálicos, todavia, a base do Nhenhém está preparada para gerar pesquisas dessa natureza, que poderão ser implementadas futuramente.

Os empréstimos lingüísticos, embora sejam menos problemáticos no caso de dados fonológicos (MALMBERG, 1993), podem causar erros de tradução, porque o valor de um fonema é dado em função dos fonemas que o ladeiam, ou seja, da combinação entre eles. Combinações atípicas são inesperadas pelo programa, por isso, devem ser editadas e controladas. Certamente, uma ou outra ocorrência dessas, em meio a dezenas de milhares de dados não prejudica a estatística gerada, mas ainda assim deve ser evitada.

A questão da representatividade – assegurar que um *corpus* efetivamente representa a língua – é mais tênue no caso de um banco de textos fonológicos, uma vez que os fonemas são em número limitado e baixo, assim, há alta taxa de repetição, que implica freqüência muito maior. O equilíbrio necessário para asseverar representatividade, mesmo parcial, é fornecido pela própria estrutura fonológica da língua, ou seja, por seus padrões, que variam menos com o tipo de conteúdo do texto analisado. Diante disso, um *corpus* fonológico não precisa ser tão grande e variado quanto um *corpus* de textos escritos, mas à medida que a massa aumenta os erros diminuem, embora não zerem.

Os relatórios de estatística de diversos textos analisados individualmente e em grupo devem apresentar valores aproximados, pois exibem os padrões fonêmicos de distribuição da língua portuguesa. Desvios dessa ordem, mesmo variações sensíveis e insistentes, em alguns casos, advertem sobre algo que merece ser descoberto e explorado. Textos muito pequenos e de tipos específicos podem não conter palavras suficientes para representar os padrões de distribuição da língua, no entanto, valores numéricos que provêm deles são legítimos.

A massa lingüística é um banco de dados. Os fonemas são a população de um *corpus* lingüístico fonológico ou fonético. A estatística faz parte das possíveis consultas (exame, pesquisa) aos dados. É mais fácil validar uma massa lingüística fonológica do que uma massa de textos escritos como representativa da língua, respeitando-se suas naturezas. O uso do sistema Nhenhém será importante para ampliar e aprimorar essa teoria.

2.4.2.1 Montagem da massa

Para testar o Nhenhém, montou-se uma massa de dados com seis artigos de odontologia, publicados em 2007, em uma revista técnico-científica da área, portanto, são textos específicos, revisados, atuais, que não foram produzidos para ser usados em pesquisa lingüística. Esses textos, depois de convertidos, totalizaram 69.787 fonemas. Na distribuição dos principais traços, esse número dividiu-se em 33.226 fonemas silábicos, 3.069 assilábicos e 33.492 consonantais, em porcentual, obtiveram-se 47,61%, 4,40% e 47,99%, respectivamente.

Antes de ser colados no Nhenhém, os textos originais foram pré-editados no Word, individualmente. De seu conteúdo eliminaram-se palavras estrangeiras – inclusive nomes próprios –, palavras que continham grafemas não pertencentes ao sistema escrito do português e unidades de medida. As siglas poderiam ser substituídas por sua forma pronunciada, bem como as unidades de medida, mas se optou por eliminá-las também (DNA, PR, SC). Sinais de pontuação e símbolos que o sistema exclui não precisam ser tratados.

Na formatação, o trabalho é maior. Foram retirados negritos, itálicos, caracteres subscritos, sobrescritos, recuos, deslocamentos, linhas em branco, tabelas, quadros, tabulações e marcadores. Os marcadores, numéricos ou com símbolos, devem ser retirados, obrigatoriamente, porque podem causar erro de leitura no Nhenhém. O espaçamento entre linhas deve ser simples. No Word 2007, há o botão 'Limpar Formatação', no menu Início, que faz todo esse trabalho automaticamente, mas, após usar esse recurso, aconselha-se a conferir o texto antes de colá-lo no programa de conversão.

Depois disso, mudou-se o idioma dos textos para português do Brasil, no menu Ferramentas (para que o editor Word checasse novamente todos os vocábulos), e verificou-se a ortografia novamente, para checar e sanar desvios remanescentes de grafia, palavras e siglas.

Para reduzir as chances de haver erro de conversão, devem-se tomar esses cuidados, que deixam o texto pronto para ser lido pelo Nhenhém. Esses seis textos foram colados no programa, convertidos, impressos, verificados, editados e salvos para pesquisa estatística. Na estatística, eles geraram, dentre outros, os números expostos no início desta seção, que, assim, são confiáveis.

2.5 ORGANIZAÇÃO INTERNA DO NHENHÉM

Na computação, há respostas lógicas para perguntas lógicas, numa combinação de linguagens. Por isso, inserir em um programa ações que parecem simples ao leigo pode envolver desenvolvimento complexo, que requer meses de pesquisa e testes. Criar programas de computador consiste em usar criatividade e pensar de forma diferente, ver as situações por diversos ângulos, prever comportamentos e interferências, montar quebra-cabeças.

Internamente, o programa contém regras de descodificação do português. Muitas vezes, no ambiente eletrônico, uma regra desdobra-se em várias outras, para dar conta das circunstâncias que ela representa. No programa, essas regras organizam-se por condições, ou seja, o sistema verbal escrito é interpretado como um complexo

conjunto de condições. Assim, na parte das regras, trabalha-se, basicamente, com a estrutura condicional **If... then... else...**

Por exemplo, é regra de descodificação da língua portuguesa que "s", quando vier entre vogais, será /z/.[30] Assim,

$$\text{"s"} \rightarrow /z/ \Big/ \text{V} \_\_ \text{V}$$

No sistema, essa regra é parte de outras que se relacionam à leitura de "s". Especificamente esse caso é internalizado da seguinte forma:

```
...
      else
      if valor[I]='s' then
      if (valor[I-1] in ['a','á','â','ã','e','i',’í’,'o','ó','ô',
      'u','é','ú','ê')]) and (valor[I+1] in ['a',’á’,’â’,'ã','e',
      'i','í','o','ó','õ','u','ú','é','ê']) then
      begin
        delete(valor,I,1);
        insert(chr($7A),valor,I);
      end;
...
```

Na regra:

- **valor** é a variável texto (cadeia de caracteres) que representa a palavra que está em teste.
- **I** é a variável numérica (inteiro) que representa a posição da letra em teste no sistema ("s" no caso).
- **in**... **and** indicam os caracteres que podem vir antes e depois de "s", para validar a regra.
- **begin**... **end** indicam o começo e o fim da operação (ou rotina) a ser realizada.
- **delete**... **insert** indicam que o valor de I (que é "s") deve ser apagado e em seu lugar, inserido /z/. É a operação.

[30] Vale comentar a grafia *autosegmental* usada por pesquisadores em fonética e fonologia. De acordo com o sistema escrito padrão da língua portuguesa, o qual deve ser usado sobretudo por educadores e pesquisadores, essa grafia gera a pronúncia [awtozegimẽtˈaw], mas o que se diz é /awtosegimẽtˈaw/, portanto, a grafia correspondente é "autossegmental".

- **chr($7A)** é o código na fonte do IPA que corresponde a /z/.

2.6 DESCODIFICAÇÃO DA ESCRITA DO PORTUGUÊS DO BRASIL

Basicamente, o código fonte do sistema de conversão grafema-fonema Nhenhém 1.0 contém a maioria das regras de descodificação da língua portuguesa, como são apresentadas por Scliar-Cabral (2003, p.81-109). Não foram utilizadas algumas regras referentes às vogais, as regras correspondentes aos ditongos e aos tritongos, bem como não foi usado o arquifonema |R| na tradução de palavras em que "r" sucede "l", "n" e "s" e em início de palavra. Já que, nessas circunstâncias, ocorre somente /r/ forte (CÂMARA JR., 1986, CAGLIARI, 2002), a letra "r" foi traduzida como fonema. Os casos de total correspondência grafofonêmica independente de contexto não foram internalizados ao sistema, simplesmente, bastou não impor condição a seu aparecimento, de modo que o programa apenas os repete na conversão.

Quanto ao restante, algumas adaptações e redefinições tiveram de ser feitas, pois se trata da conversão de um ambiente (racional, subjetivo) para outro (lógico, objetivo), além da visão da pesquisadora ser diferente, própria. A partir disso, sugerem-se alterações nas regras, na forma como são apresentadas, nos símbolos usados, por motivos coerentes e diversos, os quais são detalhados no estudo. Isso era previsível, pois a teoria de Scliar-Cabral, marco definitivo para a realização desta pesquisa, ao ser passada para a prática, como qualquer teoria, sofreu alterações e adaptações. É o curso natural da ciência, necessário para aumentar a abrangência das pesquisas.

A descodificação é muito mais simples do que a codificação (SCLIAR-CABRAL, 2003), por isso, parte-se do princípio de que o usuário tem certeza da grafia de o que informa ao programa. O sistema garante a conversão de o que é previsível na língua, para o que há regras. Isso significa que parte da língua não pode ser convertida propriamente, pois um sistema eletrônico trabalha com lógica.

No sistema Nhenhém, para os fonemas, vale o princípio "diga-me com quem andas e te direi quem és". O valor dos grafemas é determinado pelas letras que os

cercam, e não pelos fonemas, mesmo assim, essas letras nem sempre têm comportamento uniforme (como as letras “s” e “x”), o que gera desvios de regra, ou seja, exceções. Apesar disso, dada a grande transparência do sistema alfabético do português do Brasil, para a leitura, os maiores problemas são os valores abertos ou fechados dos grafemas “e” e “o” e os valores de “x” (SCLIAR-CABRAL, 2003). O desenvolvimento do sistema reafirmou isso.

Conforme mostra a revisão da literatura, há divergência na pronúncia, na separação silábica, na classificação do número de vogais e consoantes da língua portuguesa falada no Brasil. Se é assim na teoria e na prática da teoria, em um sistema eletrônico, essas disparidades ficam mais evidentes e implicam criação de novas regras e restrições de uso do programa.

2.6.1 Padronizações

A eficiência e a laboriosidade de um sistema eletrônico dependem de seu planejamento. Num sistema computacional, como em todo sistema planejado, há convenções, a fim de facilitar sua criação e padronizar seu funcionamento. No caso, seguem-se as regras fonológicas existentes até onde for possível e, então, adaptam-se e escolhem-se outras ou, para dar conta dos desvios, criam-se novas regras, dentre as opções aceitáveis cientificamente. O sistema Nhenhém 1.0 trabalha com a ortografia oficial da língua portuguesa do Brasil, vigente em 2008.

Enquanto um sistema desenvolve-se, ele exige algumas padronizações e mostra certas incoerências, que são, de fato, incompatibilidades entre sistemas de naturezas distintas ou até, no mesmo sistema. Como o aplicativo trabalha com lógica, em alguns casos de previsibilidade parcial, há como criar uma regra lógica que seja uma sub-regra de outra, assim, uma parte fica previsível e, então, é convertida pelo sistema.

À medida que os problemas surgem, mais teoria é consultada, a fim de encontrar explicação e documentação para a dificuldade e resolvê-la. Por exemplo, teve de ser feito um grupo de regras para sanar erros provocados por outras regras.

Como resultado disso, desencadeia-se um processo prático e valioso de investigação do funcionamento da língua portuguesa.

A maioria das regras de descodificação independentes do contexto foi inserida no sistema, com poucas exceções, em que se considerava o nível da sílaba, que implicavam erro em outros vocábulos, que admitiam outra definição válida e mais harmonizável com outras regras. A partir disso, não se usa o arquifonema |W|, em seu lugar, dependendo do caso, está a vogal /u/, ou /o/, ou a semivogal /w/.

Considera-se hiato a maioria dos chamados ditongos crescentes, à exceção de alguns que sucedem consoantes velares. Da mesma forma, tritongos são somente os inseparáveis, o que os limita aos que sucedem consoante velar, os outros casos configuram ditongos. Assim, o princípio adotado é o de que o ditongo, bem como o tritongo, é inseparável. Se houver separação, trata-se de hiato, portanto.

Em alguns casos, o sistema preserva uma possibilidade de pronúncia, mesmo que não seja generalizada. Algumas variantes sociolingüísticas não foram consideradas, como a transposição /w/ → /r/: “barde”, por não serem oficiais. O nível da frase (juntura vocabular, reanálise silábica) não é contemplado na versão 1.0 do Nhenhém, mas pode ser resolvido pelo usuário na edição do texto convertido.

O enfoque do aplicativo é no fonema e nos fenômenos que o rodeiam, mas inicialmente se pretendia que sua organização se pautasse na sílaba. O fonema revelou-se mais fácil para trabalhar, pois a organização silábica do português extrapola os moldes propostos, e é mais simples e rápido para o processamento eletrônico comparar os caracteres anteriores e posteriores a outro. Outrossim, testes preliminares mostraram que a irregularidade silábica seria um entrave para a execução do programa, que não poderia ser resolvido no tempo normal de pesquisa. Desse modo, majoritariamente, o programa é regido por fonemas e, raramente, morfemas, mas não pela sílaba, apesar de conceitos sobre a sílaba ser levados em conta no desenvolvimento das regras.

Por ser basicamente fonológico, o programa tenta abarcar uma pronúncia geral, válida para todos os falantes do português brasileiro. Todavia, algumas decisões em prol de certas variedades podem ser privilegiadas, dado o comportamento da linguagem de programação e o funcionamento das regras em conjunto no sistema, para

evitar conflitos entre elas. Ainda há o fator humano, nesse caso, a influência das concepções lingüísticas do pesquisador, seu idioleto.

O sistema Nhenhém parte do princípio de que o usuário está certo da grafia do texto que digita como entrada.[31] Se o usuário erra, o sistema descodifica a palavra da mesma forma, com o erro que ela contém. Fornece, portanto, a transcrição de como ela seria se aquela escrita fosse correta.

Outras padronizações, bem como detalhamentos de alguns desses pontos, estão discriminadas nas seções seguintes.

2.6.2 Letras e Fonemas

Sistematizar as consoantes é mais fácil do que sistematizar as vogais, apesar da polivalência da letra "x", que foi o maior entrave nesse sentido. Em casos não previsíveis desse grafema, foi usado o símbolo do IPA /X/, para indicar que a pronúncia do grafema "x" não pode ser automática, de forma que o usuário deve editá-lo para acertá-la. Não somente as consoantes recebem leituras diferentes (a exemplo dos valores de "x" e de "s"), isso ocorre também com vogais.

Algumas definições específicas referentes ao tratamento das vogais foram necessárias. Seja nasal ou nasalizada a vogal aparece com o mesmo diacrítico, tendo em vista que o til é o único usado com essas finalidades nas transcrições em português.

A nasalidade fonológica é sempre marcada, e a nasalidade que se dá por assimilação do traço de /m/ ou /n/ da sílaba seguinte somente é marcada quando ocorre com a vogal /a/ tônica em paroxítonas, como em "gama" → /gˈãma/, quando é inconfundível. A nasalização das outras vogais fica discreta, porque se confunde com a preparação vocal para emitir a consoante nasal seguinte. A única circunstância em que a nasalização por assimilação é marcada em todas as vogais é a mais notória, ou seja, quando antecedem /ɲ/, como em "sinhá" → /sĩɲˈa/.

[31] Antes de testar as regras, aconselha-se ler a descrição do funcionamento do programa, que está em seção posterior.

Como não é possível pronunciar certas combinações de consoantes, insere-se uma vogal entre elas, geralmente /i/. Essa epêntese ocorre também na pronúncia de palavras estrangeiras que terminam por consoante que geralmente recebem vogal em português, como "chipe" (*chip*), "nocaute" (*knock out*), quando então se torna silábica. No interior da palavra essa vogal auxiliar é similar à vogal /i/. A sistematização das regras no programa permite defender que essa epêntese não apresenta valor de vogal, pois não sonoriza o /s/ posterior a ela, como o fazem as vogais nessa mesma circunstância, pois "s" entre vogais tem valor de /z/. No entanto, ela ocorre em posição pré ou postônica, que não são circunstâncias ideais para caracterizar os fonemas vocálicos.

Realmente, esse fenômeno epentético torna-se peculiar se a letra "s" está próxima, quando a descodificação de "s" configura uma das maiores dificuldades para o programa. Se "s" entre vogais tem som de /z/ em português, quando "s" vem depois de "b" ("abster", "observar"), "d" ("adstringente"), devia ser pronunciada como /s/. Como essa combinação é atípica em português, nesses casos, a tendência é ocorrer a epêntese de uma vogal reduzida após a consoante /b/, como /subisˈidiu/. Desse modo, o "s" ficaria entre duas vogais, portanto, teria som de /z/, [subizˈidiu], mas não ocorre isso, ou seja, há casos em que esse "s" é surdo, como em "subseção", /subisesˈãw/. Portanto, não há lógica fonêmica aparente nessas ocorrências, o que incita desenvolver estudos relacionados a esse tema.

Os arquifonemas usados no programa são |S| e |R|, e ambos podem gerar conflito com outras letras durante a conversão, em alguns computadores, embora isso seja pouco provável. O primeiro corresponde às realizações de: "s" e "z" em final de palavra; "s" em final de sílaba que antecede consoante surda ou sonora e copia dela esse traço; "x" após "e" e antes de consoante surda; e "x" no prefixo "ex". O arquifonema |R| corresponde às realizações de "r" em final de sílaba.

Há letras que não têm valor fonológico, como a letra "h" em início e em final de vocábulo, portanto, não aparecem na transcrição.

O aplicativo interpreta que há em português, em encontros vocálicos, ditongos decrescentes e os ditongos crescentes “ua”, “uo”, “üe” e “üi”, em casos especiais. Assim, os ditongos do português são todos os decrescentes oficiais e os crescentes “üe” e “üi”, antecedidos por “g”; “ua” e “uo”, antecedidos por “q”; e “ua” antecedido por “g”. Esse último foi assim considerado unicamente pela pronúncia inseparável desse par, nesse tipo de sílaba, o que não ocorre com “guo”, que se separa nas conjugações de verbos terminados em “guar”: “apaziguo” → /apazigˈuu/.

Apesar de os ditongos, na escrita, serem reservados às vogais, fonologicamente, há fenômenos que igualam consoantes a vogais e geram ditongos decrescentes por neutralização de “l” em favor de /w/. É o caso de “al” (“caldo” → /kˈawdu/), que se iguala a “au” (“aura” → /ˈawra/); “el” (“tropel” → /tropˈɛw/), que se iguala a “éu” (“troféu” → /trofˈɛw/); “il” (“brasil” → /brazˈiw/) e “iu” (“corrigiu” → /koʀ̥iʒˈiw/); “ol” (“cerol” → /serˈɔw/), que em fim de vocábulo não tem oposição com vogal, por não haver “óu” em português, mas no interior de vocábulo iguala-se a “ou” (“couraça” → /kowrˈasa/, “emoldurar” → /emowdurˈaR/); e “ul” (“resultado” → /ʀ̥ezuwtˈadu/), que se assemelha ao hiato final de vocábulos terminados em “uo” (“recuo” → /ʀ̥ekˈuu/, “supérfluo” → /supˈɛRfluu/), mas é inseparável (“azul”, “sul” → /azˈuw/, /suw/).

Então, para o programa, palavras como “vaidade” e “saudade” contêm ditongo, e não hiato, uma vez que se pronunciam as duas vogais num único esforço expiratório, a exemplo de “eunuco”. O ditongo, frisa-se, é o encontro inseparável de duas vogais ou de uma vogal seguida de consoante que se vocaliza na pronúncia (“l”, “m” e “n”), portanto, de vogal com semivogal, ou, mais raramente, semivogal com vogal, na descodificação fonológica.

A parte da língua portuguesa do Brasil que o Nhenhém menos cobre é a harmonia vocálica e o timbre aberto não marcado das vogais /e/ e /o/. Por isso, antes de se fazer pesquisa sobre as vogais, o texto deve ser ajustado na edição. Sistematizar esses casos, em algumas circunstâncias, é possível, mas o nível de complexidade na programação é elevado, como nos casos de alguns verbos. Fazer

listas de exceções para essas situações tornaria o sistema menos lógico. Seria uma solução pouco inteligente. Nesse sentido, o programa considera apenas palavras terminadas em “osa”, “osos”, “el” e “ol”, “el” tônico no interior de vocábulo e as que recebem acento gráfico.

O Nhenhém não interfere na pronúncia de “coalho” como /kuˈaʎu/ e “alardear” como /alaRdiˈaR/, porque isso tornaria errada a tradução de “boa”, “reagir” e “tear”, por exemplo. Então, para o programa as traduções dessas palavras são: /koˈaʎu/ e /alaRdeˈaR/. Testes preliminares revelaram que há como inserir regras para a pronúncia de “e” e “o” nesses casos, mas são regras elaboradas, com programação complexa, cujo desenvolvimento demanda pesquisa e tempo.

2.6.3 Tonicidade

Não foi tão complexo inserir a prosódia no sistema, porque ela leva em conta a terminação da palavra, na grande maioria dos casos. Como o sistema de acentuação gráfica da língua portuguesa é lógico, a ortoépia, que inicialmente parecia difícil de informatizar, surpreendeu. Com a adaptação da teoria lingüística para a teoria de computação e redefinição de alguns pontos divergentes da teoria lingüística, o Nhenhém 1.0 ficou preparado para lidar com a acentuação de qualquer palavra da língua portuguesa, com raríssimas exceções, estima-se, pois em todos os testes feitos não ocorreu erro de prosódia.

Uma das definições necessárias, nesse aspecto, foi com relação ao local onde apareceria a marca fonológica de tonicidade /ˈ/. Se, como ensina a teoria, o pico da sílaba é a vogal, então, quando a sílaba for tônica, a vogal dela pode receber o acento, em vez da primeira letra dessa sílaba. Essa decisão tornou o sistema apto a procurar a vogal tônica a partir da terminação do vocábulo.

O que é acentuado graficamente não precisa de regra no programa, pois o usuário insere o acento, à medida que digita ou cola o texto a ser traduzido. O que não é acentuado graficamente deve ser acentuado pelo Nhenhém. Assim, o sistema verifica as palavras que não são acentuadas graficamente e a elas aplica as regras de acentuação

não marcada da língua portuguesa, antes de fazer a tradução fonológica, porque a acentuação é regida pela escrita. Por exemplo, a palavra “amam”, fonologicamente, termina em ditongo, mas não na escrita. Então, ela recebe acento na penúltima vogal, que está na penúltima sílaba, posto que termina em “am”, e não em “ão”. Desse modo, tem-se em “mamão” e “mamam” duas palavras inconfundíveis, exclusivamente por causa da prosódia: /mamˈãw/ e /mˈãmãw/, assim como “venceram” e vencerão” /vẽsˈerãw/ e /vẽserˈãw/.

Em conformidade com isso, aplicam-se regras de acentuação a palavras oxítonas e paroxítonas cuja tonicidade não é marcada graficamente. Nesse caso, a terminação da palavra define o acento. Então, recebem acento na última vogal as palavras terminadas por “r”, “l”, “x”, “z”, “n”, “um”, “im”, “om” e “ã”, bem como os vocábulos terminados por “i” e “u”, por ditongo nasal e oral e por tritongo. Nesse sentido, a maior dificuldade foi sistematizar a prosódia em palavras terminadas por “que”, “qui”, “gue” e “gui”.

Recebem acento na penúltima vogal os vocábulos terminados por “a”, “e”, “o”, “am”, “em”, “ens”. O acréscimo da letra “s” não invalida as regras. As vogais “i” e “u” em posição de semivogal não podem receber acento.

Os clíticos desviam-se das regras do português, pois, se há vogais neles, eles deveriam ter tonicidade, sobretudo se há mais de uma vogal. A solução foi pô-los numa biblioteca, uma lista de palavras que são exceções da língua (“para”, “porque”). O sistema marca o acento de palavras com mais de uma vogal, então, apenas os clíticos com mais de uma vogal estão na biblioteca. Dessa lista também fazem parte palavras com pronúncias irregulares, como “exu”,[32] “muito” e derivadas e vocábulos com o radical “troux”.

O hífen é excluído do texto na conversão, assim, palavras unidas por hífen aparecem separadas, com duas prosódias, se for o caso.

[32] A pesquisa mostra que essa palavra deveria ser escrita “echu” para adaptar-se aos padrões da língua portuguesa.

2.6.4 Algumas Limitações

Além das vogais abertas não acentuadas graficamente, há outras limitações do Nhenhém em sua primeira versão. A pronúncia das formas verbais terminadas em “ar” e “er”, como “falar”, “comer”, dispensa o “r” final na pronúncia de diversas variedades sociolingüísticas. Da mesma forma, o “s” final das formas verbais desaparece na fala, como em “vamos” → [vˈamu]. Assim, “vamos ver” pronuncia-se [vˈamu ve]. O sistema não distingue formas verbais, e sempre traduz esses “r” e “s” como pronunciados. Isso não configura erro, mas sim uma possibilidade de pronúncia que pode ser considerada padrão, aliás, usada em conversas formais e na televisão. Certamente, o usuário pode adaptar a pronúncia na edição do texto convertido, se for o caso.

Então, em algumas situações, o programa faz descodificação em uma pronúncia que nem sempre é a mais comum, mas é possível e também é usada, de forma que um usuário que a usa não é tomado por estrangeiro. Isso ocorre sobretudo com alguns encontros vocálicos.

Os morfemas “ob” e “sub” seguidos por “s” e “trans” seguido por vogal são imprevisíveis quanto ao valor fonológico de “s”, portanto, geram erro nos casos em que esse “s” deve ser descodificado com /z/. Por conta disso, a descodificação do radical “trans”, como em “transação”, fica prejudicada.

O sistema tem uma lista de sinais de pontuação e símbolos que devem ser ignorados. Os sinais de pontuação mais comuns em textos estão inseridos nessa biblioteca, porém, os símbolos são inúmeros, de forma que alguns símbolos raros podem aparecer no texto e gerar um caractere estranho na tradução.

Também há regras que o sistema não contempla, por dependerem de metalinguagem ou do contexto morfossintático e semântico, ou seja, casos imprevisíveis por regras lógicas, como os timbres abertos de /e/ e /o/ em certos verbos rizotônicos, em certos tempos verbais. Essas situações, portanto, geram necessidade de o usuário editar o texto convertido. No sistema, também pode haver regras que provocam erros advindos de conflitos com outras regras e regras que geram alguns

desvios, mas ainda não apareceram nos testes. O uso informará quais são, para que sejam resolvidos, afinal, esta é a versão 1.0, e a partir de agora será disponibilizada ao público interessado.

2.7 PERCURSO E PERCALÇOS

Naturalmente, a própria construção do sistema, sobretudo em seus entraves e percalços, é reveladora de comportamentos da língua. Por converter e analisar textos corridos e longos, em seu percurso, o programa, cujo funcionamento é resultado da interação linguagem de programação-configurações do computador do usuário-língua portuguesa do Brasil, apresentou comportamentos peculiares. Eis alguns, dentre tantos.

O fonema correspondente a “r” em início de palavra, “rr” e “r” após “l”, “n” e “s” representado por /R̥/ no Nhenhém, porque houve conflito com o arquifonema |R|, uma vez que ambos usam o mesmo caractere, porém, o fonema tem tamanho menor. Em IPA, o caractere correspondente ao fonema seria o 0091, que não é lido corretamente pela linguagem de programação Delphi, como mencionado. Reduzir o tamanho de fonte não resolveria o problema, pois o código hexadecimal continuaria o mesmo. A solução adotada baseia-se no IPA, em que o símbolo /◌̥/ indica que o fonema que o contém é surdo, e isso confere com a característica do fonema em que esse símbolo é usado no programa. Ainda, a presença de um símbolo similar ao fonema original faz o sistema ser mais intuitivo ao usuário. Até chegar-se a essa solução, muitos testes foram feitos com outros símbolos e combinações, e sempre havia incompatibilidade entre os ambientes envolvidos ou inconsistência no programa.

Indubitavelmente, a maior dificuldade foi sistematizar as vogais, sobretudo quando havia encontro vocálico. O /l/ descodificado como /w/ conflitava com ditongos terminados em /w/, como em “lousa” e “bolsa”. Nesses casos, as transcrições são: /lˈowza/ e /bˈowsa/. Pela lógica, o valor da letra “s” depois do ditongo deveria ser o mesmo, no entanto, não é. Inserir essa diferença no sistema demandou certo tempo.

Outro entrave que ocorreu no programa também se relaciona com a semivogal /w/, seguida pelas consoantes /r/ brando e forte, como em “aura” e “melro”. A

transcrição dessas duas palavras é /ˈawra/ e /mˈɛwɾ̥u/. Uma vez que um fonema sofre interferência do fonema anterior, novamente, como inserir no sistema que, apesar de se tratar, nos dois casos, da semivogal /w/, a letra posterior a ela, embora se escreva da mesma forma, configura dois fonemas distintos? A solução foi internalizar uma regra específica para esse caso, ou seja, uma sub-regra.

Quanto à prosódia, configurou o maior impasse inserir a tonicidade de hiatos "i" que antecedem "nh". Esse é o único caso em que o hiato tônico /i/ não recebe acento gráfico, então, a lógica é lê-lo como semivogal. Vale comparar três casos: "polaina", "tubaína", "ladainha" → /polˈayna/, /tubaˈina/, /ladaˈɪ̯na/. A solução foi desenvolver uma regra para esse único caso, ou seja, quando há a seqüência vogal+inh*.

Complexo também foi lidar com os encontros vocálicos não pronunciados e pronunciados com "gu", "gú" e "qu" seguidos de "e", "i", "o" e "a", com e sem acento gráfico. Nesse caso, o problema foi com a ortoépia, pois a tendência lógica é descodificá-los como ditongo. Quando se trata de "u" não pronunciado, mas escrito, o sistema deve de ignorá-lo, para não acentuar a vogal errada. O acerto dessa regra prejudicava outras, e o mais prejudicado era o "gu" pronunciado, como em "algum", "figura" e "segundo".

Uma das certezas da língua portuguesa é que a terminação "osa" lê-se como /ɔza/, então, essa deveria ser uma conversão fácil para o sistema. No entanto, como o programa relê as regras ciclicamente (pois repete a execução de rotinas até que não haja mais caracteres a analisar), transformaria a terminação "ossa" em /ɔza/, porque, na primeira comparação de regras, ambas as terminações ficam iguais: "ossa" → /osa/, então, na segunda verificação (que seria a primeira se a terminação inicial fosse "osa"), ocorre o erro /osa/ → /ɔza/. Para resolver esse obstáculo, foi necessário fazer uma regra específica para releitura da palavra. A solução desse problema requereu quinze dias. Com essa experiência, a inserção de /ɔzuS/ ficou simples, pois essa regra é controlada pelo plural das palavras terminadas em /ozu/, que são os masculinos plurais das palavras terminadas em /ɔza/.

2.8 ALFABETO FONOLÓGICO

Internamente no Nhenhém, o alfabeto fonológico da língua portuguesa é composto pelos seguintes símbolos, com os seguintes traços característicos, criado com base na Nomenclatura Gramatical Brasileira (NGB, 1959), Scliar-Cabral (2003), Bechara (1973), Câmara Jr. (1997; 1986) e em intuições próprias.

Quadro 4: Sistema de vogais fonológico do português brasileiro

	Traços	**/a/**	**/e/**	**/ɛ/**	**/i/**	**/o/**	**/ɔ/**	**/u/**	**/y/**	**/w/**	**/ã/**	**/ẽ/**	**/ĩ/**	**/õ/**	**/ũ/**	**/ỹ/**
Função na sílaba	silábico	x	x	x	x	x	x	x			x	x	x	x	x	
	assilábico								x	x						x
Via de emissão	oral	x	x	x	x	x	x	x	x	x						x
	nasal										x	x	x	x	x	
Zona de articulação	anterior		x	x	x				x			x	x			x
	posterior	x				x	x	x		x	x			x	x	
Timbre	alto				x			x	x	x			x		x	x
	médio		x			x						x		x		
	baixo	x		x			x				x					
Movimento dos lábios	arredondado					x	x	x		x				x	x	
	distenso	x	x	x	x				x		x	x	x			x

Ao Quadro 4 corresponde a seguinte legenda:

- **Silábico**: pode ser ápice de sílaba;
- **Assilábico**: não pode ser ápice de sílaba;
- **Oral**: produzido com o véu palatino levantado, de modo que a corrente de ar atravessa a cavidade oral;
- **Nasal**: produzido com o véu palatino abaixado, de modo que parte da corrente de ar atravessa a cavidade bucal, e parte, a cavidade nasal;
- **Anterior**: fonema para cuja emissão a língua se eleva em direção ao palato duro;
- **Posterior**: produzido com a língua em elevação na direção do véu palatino (velar);
- **Alto**: produzido com alto grau de elevação da língua (fechado);
- **Médio**: articulado com grau médio de elevação da língua;

- **Baixo**: produzido sem elevação da língua (aberto);
- **Arredondado**: produzido com arredondamento dos lábios;
- **Distenso**: produzido sem arredondamento dos lábios.

Todas as vogais contêm o traço de sonoridade.

Quadro 5: Sistema de consoantes fonológico do português brasileiro

	Traços		/p/	/b/	/f/	/v/	/m/	/n/	/d/	/t/	/s/	/z/	/l/	/ʎ/	**/r/**	/ɾ̥/	/ʃ/	/ʒ/	/ɲ/	/k/	/g/	|S|	|R|
Função na sílaba	consonantal		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Modo de articulação	oclusivo		x	x			x	x	x	x									x	x	x		
	constritivo	fricativo			x	x					x	x					x	x				x	
		lateral											x	x									
		vibrante													x	x							x
Ponto de articulação	labial		x	x	x	x	x																
	anterior							x	x	x	x	x	x		x							x	
	posterior													x		x	x	x	x	x	x		
Via de emissão	oral		x	x	x	x			x	x	x	x	x	x	x	x	x	x		x	x	x	x
	nasal						x	x											x				
Fonação	surdo		x		x					x	x					x	x			x			
	sonoro			x		x	x	x	x			x	x	x	x			x	x		x		

- **Oclusivo** ou plosivo: fonema consonantal produzido com fechamento total à passagem do ar;
- **Constritivo**: fonema consonantal em cuja emissão os articuladores ficam muito próximos, de modo que se forma ruído, mas há passagem parcial da corrente de ar. Pode ser fricativo, lateral e vibrante:
  - Fricativo: ao passar, a corrente de ar produz atrito por fricção;
  - Lateral: o ar escapa pelos dois lados da língua;
  - Vibrante: o ápice da língua executa movimento vibratório rápido, abre e fecha a passagem à corrente de ar;
- **Labial**: aquele em que um ou ambos os lábios, mas não a língua, participam da articulação (bilabiais e labiodentais);
- **Anterior**: aquele em que a língua se eleva em direção aos dentes incisivos superiores, aos alvéolos ou ao pré-palato;

- **Posterior**: aquele em que a parte posterior da língua se eleva em direção ao fundo da boca, do palato mole para trás;
- **Nasal**: aquele em que, simultaneamente a uma obstrução à passagem do ar no trato oral, ocorre o abaixamento do véu palatino, permitindo que parte do ar escoe pela cavidade nasal;
- **Sonoro**: aquele em que há vibração das pregas vocais;
- **Surdo**: aquele em que não há vibração das pregas vocais.

2.9 REGRAS DO NHENHÉM

Não há regras elaboradas para a grafia[33] de “p”, “b”, “f”, “v”, “ss”, “ç”, “sç”, “ch”, “j”, “rr”, “ü”, “ó”, “á”, “à”, “â”, “ã”, “õ”. Eles correspondem, respectivamente, aos caracteres: /p/, /b/, /f/, /v/, /s/, /s/, /s/, /ʃ/, /ʒ/, /ʀ̥/, /w/, /ˈɔ/, /ˈa/, /ă/, /ˈã/, /ã/, /õ/. O restante depende de regras mais complexas.

O sistema Nhenhém responde da seguinte forma às entradas:

- “p” é descodificado como /p/ (“par”→ /paR/).
- “b” é descodificado como /b/ (“bar” → /baR/).
- “f” é descodificado como /f/ (“fim” → /fĩ/).
- “v” é descodificado como /v/ (“vir” → /viR/).
- “ss” é descodificado como /s/ (“assar” → /asˈaR/).
- “ç” é descodificado como /s/ (“louça” → /lˈowsa/).
- “sç” é descodificado como /s/ (“floresça” → /florˈesa/).
- “ch” é descodificado como /ʃ/ (“choupana” → /ʃowpˈãna/).
- “j” é descodificado como /ʒ/ (“jazer” → /ʒazˈeR/).
- “t” é descodificado como /t/ (“tom” → /tõ/).[34]
- “d” é descodificado como /d/ (“dom” → /dõ/).
- “nh” é descodificado como /n/ se ocorrer antes de “ia” (“companhia” → /kõpãnˈia/).
- “nh” é descodificado como /ɲ/ nos demais contextos (“aranha” → /arˈãɲa/).
- “a”, “e”, “i”, “o” e “u” nasalizam-se se ocorrerem antes de “nh” (“unha” → /ˈũɲa/).

[33] Usa-se a palavra grafia porque o sistema trabalha com dados digitados, e não com conversão de voz.
[34] Na maioria das variações sociolingüísticas do Brasil, “t” é descodificado como [tʃ] se ocorrer antes de “i”, e “d” é descodificado como [dʃ] se ocorrer antes de “i”, ou seja, tornam-se africadas.

- “a” é descodificado como /ã/ em final de sílaba tônica que antecede “m” ou “n” que inicia sílaba seguinte (“estamos”, “fama”, “cano” → /eStˈãmuS/, /fˈãma/, /kˈãnu/).
- “ü” é descodificado como /w/ (“agüei” → /agwˈey/).
- “ó” é descodificado como /ˈɔ/ (“pó” → /pˈɔ/).
- “á” é descodificado como /ˈa/ (“pá” → /pˈa/).
- “à” é descodificado como /ă/ (“àquilo” → /ăkˈilu/).
- “â” é descodificado como /ˈã/ (“lânguida” → /lˈãgida/).
- “ã” é descodificado como /ã/ (“galã” → /galˈã/).
- “aa” é descodificado como /a/ (“caatinga” → /katˈiga/).
- “õ” é descodificado como /õ/ (“aviões” → /aviˈõyS/).
- “lh” é descodificado como /ʎ/ se ocorrer antes de “a”, “e”, “o” e “u”, no começo ou interior de vocábulos (“olheiro”, “lhufas” → /oʎˈeyru/, /ʎˈufaS/).
- “lh” é descodificado como /l/ se ocorrer antes de “i” (“canalhice” → /kanalˈisi/).
- “lh” é descodificado como /l/ se ocorrer antes de “e” em final de vocábulo (“entalhe” → /ẽtˈali/).
- “h” é descodificado como nulo se ocorrer em início e em final de vocábulo (“humanidade” → /umanidˈadi/).
- “s” é descodificado como /z/ quando ocorrer entre vogais[35] (“coisa” → /kˈoyza/).
- “s” é descodificado como |S| em fim de vocábulo (“hélices” → /ˈɛlisiS/).

[35] As vogais englobam as não acentuadas, as acentuadas graficamente e as consideradas semivogais, pois a descodificação é feita a partir da escrita, que não tem semivogais. As vogais acentuadas graficamente não são especificadas, porque são as mesmas, porém, algumas vezes, com variação de timbre. Casos particulares em que o acento gráfico interfere na regra são indicados.

- “s” é descodificado como |S| se ocorrer antes de consoante (“casta” → /kˈaSta/, “prisma” → /prˈiSma/).[36]
- “s” é lido como /s/ nos demais contextos (início de vocábulo, depois de “n”, “l”, “r”) (“sopapo” → /sopˈapu/, “denso” → /dˈẽsu/, “balsa” → /bˈawsa/, “verso” → /vˈɛRsu/.[37]
- “c” é descodificado como /s/ se ocorrer antes de “e” e “i”, nos demais contextos, “c” é descodificado como /k/ (“cegonha” → /segˈõɲa/, “caco” → /kˈaku/).
- “sc” é descodificado como /s/ se ocorrer antes de “e” e “i” (“florescer” → /floresˈeR/, “descida” → /desˈida/).
- “xc” é descodificado como /s/ se ocorrer antes de “e” e “i” (“excelência” → /eselˈẽsia/).
- “z” é descodificado como |S| em final de vocábulo (“reluz” → /ʀ̥elˈuS/).[38]
- “x” é descodificado como /ʃ/ se ocorrer após “en” (“enxada” → /ẽʃˈada/).
- “x” é descodificado como /ʃ/ se ocorrer em início de vocábulo (“xisto” → /ʃˈiStu/).
- “x” é descodificado como /ʃ/ se ocorrer após os ditongos “ai”, “ei” e “ou” (“faixa” → /fˈayʃa/, “ameixa” → /amˈeyʃa/, “trouxa” → /trˈowʃa/).
- O radical “troux” é descodificado como /trˈows/ (“trouxe” → /trˈowsi/).

---

[36] Trata-se da assimilação que ocorre com o /s/ que precede consoante sonora, e copia dela o traço sonoro, transformando-se em /z/, ou permanecendo como /s/, se a consoante seguinte for surda: /mˈista/ e /sˈizma/.

[37] Pela lógica do programa, o “s” do prefixo “trans” seria sempre lido como /s/, antes de fonema surdo (transporte, trans+sexual – *transsexual*) e como /z/ antes de fonema sonoro (trans+amazônica – “tranzamazônica”), no entanto, há casos (“transação” = do latim *transactione*; “transecular” = trans+secular – *transsecular*) em que a pronúncia depende da origem da palavra, mas a forma canônica é a pronúncia /s/ (AURÉLIO, 2004). Assim o programa gera o erro [trãsasˈãw]. Também configura irregularidade o som de /z/ após /b/ (“obséquio”, “observar”, “subseção”, “subsídio”).

[38] Trata-se do arquifonema |S|, no caso em que “s” e “z” neutralizam-se, quando ocorrem em fim de vocábulo, antes de silêncio.

- “x” é descodificado como |S| se ocorrer entre “e” em início de vocábulo e “c” duro, “f”, “p”, “t” (“excluir”, “explicar” → /eSkluˈiR/, /eSplikˈaR/).
- “x” é descodificado como /kis/ se ocorrer em fim de palavra (“sílex” → /sˈilekis/).
- “x” é descodificado como /z/ se ocorrer entre “e” em início de vocábulo e “a”, “e”, “o” e “u” (“execução” → /ezekusˈãw/).
- “x” é descodificado como /X/, nos casos em que não pode ser previsto por regras (“coxilha” → /koXˈiʎa/).
- “g” é descodificado como /ʒ/ se ocorrer antes de “e” e “i” (“agente” → /aʒˈẽti/).
- “gu” é descodificado como /g/ quando ocorrer antes de “e” e “i” (“gueixa” → /gˈeyʃa/).
- “qu” é descodificado como /k/ se ocorrer antes de “e” e “i” (“química” → /kˈimika/).
- “qu” é descodificado como /kw/ se ocorrer antes de “a” e “o” (“quântico” → /kwˈãtiku/).
- “q” é descodificado como /k/ se ocorrer antes de “ü” (“liqüidificador” → /likwidifikadˈoR/).
- “g” é descodificado como /g/ se ocorrer antes de “a”, “o”, “u”, “r”, “l” (“gatuno”, “pagã”, “sagüi”, “agrura”, “glacê” → /gatˈunu/, /pagˈã/, /sagwˈi/, /agrˈura/, /glasˈe).
- “m” é descodificado como /m/ se ocorrer antes de vogal (“margarida” → /maRgarˈida/).
- “n” é descodificado como /n/ se ocorrer antes de vogal (“namoro” → /namˈoru/).
- “m” é descodificado como /y/ se ocorrer após “e”, em final de vocábulo (“convém”, “vêm”, “contem” → /kõvˈẽy/, /vˈẽy/, /kˈõtẽy/).

- “n” é descodificado como /y/ se ocorrer após “e”, em final de vocábulo (“hímen” → /ˈimẽy/).
- “e”, “é”, “ê” nasalizam-se se ocorrerem antes de “m” em final de vocábulo (“comem” → /kˈomẽy/).
- “e”, “é” nasalizam-se se ocorrerem antes de “n” em final de vocábulo (“lúmen” → /lˈumẽy/).
- “m” e “n” nasalizam a vogal que as precede, se ocorrerem em final de sílaba interna (“dormente”, “bomba” → /doRmˈẽti/, /bˈõba/).
- “m” em final de vocábulo nasaliza as vogais precedentes “i”, “o” e “u” (“curumim”, “bombom”, “fórum” → /kurumˈĩ/, /bõbˈõ/, /fˈɔrũ/).
- “m” em final de vocábulo nasaliza a vogal precedente “a” e semivocaliza-se em /w/ (“facilitam” → /fasilˈitãw/).[39]
- “n” em final de vocábulo nasaliza as vogais “i”, “o” e “u” (“sarin”, “cólon”, “bodun” → /sarˈĩ/, /kˈɔlõ/, /bodˈũ/).[40]
- “l” é descodificado como /w/ se ocorrer em final de sílaba, ou seja, quando a ele não suceder vogal (“cálcio” → /kˈawsiu/).[41]
- “el” é descodificado como /ɛw/ se ocorrer em fim de vocábulo oxítono (“coronel” → /koronˈɛw/).

[39] Nesse caso, pelas regras ortográficas do português, a palavra é sempre paroxítona, a menos que seja uma sigla ou empréstimo lingüístico. Assim a palavra é oxítona se terminar em “ão” e paroxítona se terminar em “am”, se não forem, recebem acento gráfico. O dicionário (AURÉLIO, 2004) registra apenas siglas e palavras estrangeiras com a terminação “am”, que se iguala a /ã/, como “diazepam”, que deveria ser “diazepã”, pois a terminação “am” é restrita a alguns verbos conjugados em português.

[40] Em final de vocábulo, o que seria “an” grafa-se como “ã”, embora haja exceções em estrangeirismos, em que a falta de conhecimento do sistema escrito da língua portuguesa causa desvios. Por exemplo, ao falar da guerra no Iraque, houve meios de comunicação que adotaram a grafia correta em português “Talibã”, enquanto outros usaram *Taliban*. Adequada é a forma Talibã, assim como Irã. Depois de “o”, mesmo que a grafia contrarie as regras do português, não há problema para o programa, pois, nesse caso, não há diferença de pronúncia entre “on” e “om”. Depois de “u”, o “n” ocorre em plurais de vocábulos terminados em “um”.

[41] A variedade dita caipira, em que esse “l” torna-se “r” retroflexo não é considerada pelo sistema, bem como o linguajar de parte do Rio Grande do Sul, em que esse “l” se torna velar [ɬ].

- “ol” é descodificado como /ɔw/ se ocorrer em fim de vocábulo oxítono (“sorbitol” → /soRbitˈɔw/).
- “l” é descodificado como /l/ se preceder vogal (“calo” → /kˈalu/).
- “l” é descodificado como /l/ se suceder “b”, “c”, “f”, “g”, “p”, “v” (“flamengo”, “clique” → /flamˈẽgu/, /klˈiki/).
- “rr” é descodificado como /Ɽ/ (“borrado” → /boƦˈadu/).
- “r” é descodificado como/Ɽ/ se ocorrer em início de vocábulo (“rosa” → /ƦˈɔzaA/).
- “r” é descodificado como /Ɽ/ se ocorrer depois de “l”, “n” e “s” (“palra”, “enrosco”, “desramar” → /pˈawƦa/, /ẽƦˈoSku/, /deSƦamˈaR/).
- “r” é descodificado como |R| se ocorrer em final de sílaba (“torcer” → /toRsˈeR/).
- “r” é descodificado como /r/ se ocorrer entre vogais (“cera” → /sˈera/).
- “r” é descodificado como /r/ se ocorrer depois de “b”, “c”, “d”, “f”, “g”, “p”, “t”, “v” (“croma”, “franco”, “livrar” → /krˈoma/, /frˈãku/, /livrˈaR/).
- “ê” é descodificado como /ˈẽ/ se ocorrer antes de “m” ou “n” travando sílaba interna de vocábulo (“amêndoa” → /amˈẽdua/).
- “ê” é descodificado como /ˈe/ nos demais contextos (“pêssego” → /pˈesegu/).
- “êem” é descodificado como /ˈẽy/ (“vêem” → /vˈẽy/).
- “ôo” é descodificado como /ˈou/ (“revôo” → /Ɽevˈou/).
- “ee” é descodificado como /e/ quando for átono (“reescrever” → /ƦeSkrevˈeR/).
- “oo” é descodificado como /o/ quando for átono (“coordenar” → /koRdenˈaR/).
- “oó” é descodificado como /ˈɔ/ (“alcoólico” → /awkˈɔliku/).
- “ô” é descodificado como /ˈõ/ se ocorrer antes de “m” ou “n” travando sílaba não final de palavra (“bômbus” → /bˈõbuS/).

- “ô” é descodificado como /ˈo/ nos demais contextos (“bônus” → /bˈonuS/).
- “í” é descodificado como /ˈ̃i/ se ocorrer antes de “m” ou “n” travando sílaba (“língua” → /lˈ̃igwa/).
- “í” é descodificado como /ˈi/ nos demais contextos (“cínico” → /sˈiniku/).
- “ú” é descodificado como /ˈũ/ se ocorrer antes de “m” ou “n” travando sílaba (“plúmbeo” → /plˈũbiu/).
- “ú” é descodificado como /ˈu/ nos demais contextos (“púnico” → /pˈuniku/).
- “é” é descodificado como /ˈẽ/ se ocorrer em final de vocábulo, antes de “m” ou “n” (“vintém” → /vĩtˈẽy/).
- “é” é descodificado como /ˈɛ/ nos demais contextos (“café” → /kafˈɛ/).
- “e” é descodificado como /i/ se ocorrer em final de vocábulo, em sílaba átona (“fonte” → /fˈõti/).
- “o” é descodificado como /u/ se ocorrer em final de vocábulo, em sílaba átona (“mano” → /mˈãnu/).
- “o” é descodificado como /w/ se ocorrer depois de “ã” (“maganão” → /maganˈãw/).
- “osa” é descodificado como /ɔza/ se ocorrer em final de vocábulo (“fabulosa” → /fabulˈɔza/).
- “oso” é descodificado como /ɔzu/ se ocorrer em final de vocábulo e seguido de “s” (“fabulosos” → /fabulˈɔzuS/).
- “ea” é descodificado como /ia/ em final de vocábulo (“alínea” → /alˈinia/).
- “eo” é descodificado como /iu/ em final de vocábulo (“cetáceo” → /setˈasiu/).
- “ie” é descodificado como /ii/ em final de vocábulo (“calvície” → /kawvˈisii/).
- “io” é descodificado como /iu/ em final de vocábulo (“adágio” → /adˈaʒiu/).

- “oa” é descodificado como /ua/ em final de vocábulo proparoxítono (“névoa” → /nˈɛvua/).
- “oe” é descodificado como /oi/ em final de vocábulo (“abençoe” → /abẽsˈoi/).
- “ua” é descodificado como /wa/ quando suceder “g” ou “q” (“aguar”, “aquavia” → /agwˈaR/, /akwavˈia/).
- “ua” é descodificado como /ua/ em final de vocábulo, quando não suceder “g” ou “q” (“continua” → /kõtinˈua/).
- “ue” é descodificado como /ui/ em final de vocábulo, quando não suceder “g” ou “q” (“continue” → /kõtinˈui/).
- “uo” é descodificado como /wo/ se anteceder “q” (“quotista” → /kwotˈiSta/).
- “uo” é descodificado como /uu/ em fim de vocábulo (“amuo”, “contíguo” → /amˈuu/, /kõtˈiguu/).
- “i” é descodificado como /y/ se ocorrer depois de “a”, “e”, “o” e “u” na mesma sílaba (ditongo decrescente) (“baixada”, “quartéis”, “canseira”, “bemóis”, “moita”, “circuito” → /bayʃˈada/, /kwaRtˈɛyS/, /kãsˈeyra/, /bemˈɔyS/, /mˈoyta/, /siRkˈuytu/).
- “ui” é descodificado como /uỹ/em “muito” e derivados (/mˈuỹtu/).
- “ãe” é descodificado como /ãy/ (“cães” → /kãyS/).
- “ão” é descodificado como /ãw/ (“bênção” → /bˈẽsãw/).
- “am” é descodificado como /ãw/ em fim de vocábulo (“encorajam” → /ẽkorˈaʒãw/).
- “em” é descodificado como /ẽy/ em fim de vocábulo (“caem’ → /kˈaẽy/).
- “õe” é descodificado como /õy/ (“petiçõezinhas” → /petisõyzˈɪɲaS/).

- “u” é descodificado como /w/ se ocorrer depois de “a”, “e”, “é”, “i” e “o” na mesma sílaba (ditongo decrescente) (“autor”, “escarcéu”, “euforia”, “feriu”, “besouro” → /awtˈoR/, /eSkaRsˈɛw/, /ewforˈia/, /ferˈiw/, /bezˈowru/).
- “uai” é descodificado como /way/ quando vier após “g” ou “q”, ou seja, quando for tritongo (“guaipé”, “quais” → /gwaypˈɛ/, /kwayS/).
- “uão” e “uam” são descodificados como /wãw/ quando vierem após “g” ou “q”, ou seja, quando forem tritongo (“enxáguam”, “saguão”→ /ẽʃˈagwãw/, /sagwˈãw/).[42]
- “üei” é descodificado como /wey/ (“enxagüei” → /ẽʃagwˈey/).
- “oou” é descodificado como /uow/ em final de vocábulo (“abençoou” → /abẽsuˈow/).
- Encontros consonantais imperfeitos, isto é, quando as consoantes não pertencem à mesma sílaba, podem ser descodificados com acréscimo de /i/ auxiliar (“naftalina”, “advérbio” → /nafitalˈina/, /adivˈɛRbiu/).

2.10 ALGUMAS CONSIDERAÇÕES

No programa, há mais de 200 regras para dar conta dessas, pois várias delas desdobram-se em duas ou mais, para que a descodificação seja feita.

A descodificação do Nhenhém tenta ser coerente com a lógica do sistema verbal, que nem sempre é lógico. Assim, variações de pronúncia que dependem de sub-regras das regras lógicas podem não estar contempladas na versão 1.0 do programa, porque seu desenvolvimento e inserção no programa demandam mais testes. A inclusão de regras não suficientemente testadas pode gerar conflito com outras regras, o que se manifesta como erro na tradução e, conseqüentemente, nos relatórios.

[42] A única palavra com a terminação “uão” registrada no dicionário é essa, com sua variação “xaguão”.

O til como marcador de vogal que ocorre antes de consoante nasal tende a causar certos problemas a um sistema eletrônico descodificador, que trabalha mais bem se interpretar cada som como um caractere apenas ou dois caracteres de leitura internacional. Assim, mais fácil seria trabalhar com /aN/ em vez de /ã/, por exemplo, mas, após algumas tentativas, obteve-se consistência com uso de til, pois pode haver outros grafemas que são representados fonologicamente por dois símbolos – como /ʀ̥/ –, de modo que é viável deixar o sistema preparado para lidar com casos similares a esses.

Para bem atender às necessidades do português do Brasil e não gerar conflito com outros fonemas, o IPA deveria ter o caractere /ʃ/ para ser usado em [tʃ]. O ganho seria também estético, mas esse símbolo não é usado pelo Nhenhém, portanto, é apenas uma sugestão. Os símbolos que o IPA disponibiliza para tradução da consoante "r" não atendem à necessidade da língua portuguesa, porque geram incompatibilidade e são ilegíveis por algumas linguagens de programação.

Todo sistema eletrônico é mais eficiente nas mãos de usuários eficientes. Se o usuário conhece as limitações e as virtudes do programa, se o usa de acordo com a proposta do programa, se segue as regras que regem seu funcionamento, o desempenho de ambos será melhor. Um usuário criativo pode ampliar os recursos do programa, explorá-los, descobrir diferentes formas de usá-los a favor de sua pesquisa.

3 DESCRIÇÃO DO USO DO PROGRAMA NHENHÉM 1.0

o bicho alfabeto
tem vinte e três patas
ou quase
por onde ele passa
nascem palavras e
frases
com frases
se fazem asas
palavras
o vento leve
o bicho alfabeto passa
fica o que não se escreve

Leminski (1991)

Figura 4: Tela de abertura do sistema Nhenhém

3.1 INTRODUÇÃO

Conhecer bem o funcionamento do sistema eletrônico de conversão grafema-fonema Nhenhém© 1.0 é condição para a interação usuário-sistema. Desse modo, o usuário saberá o que esperar do programa, como tratar desvios de conversão e, sobretudo, usá-lo de maneira eficiente. As etapas principais do sistema são: conversão (tradução automática do texto para o alfabeto fonético), edição (manipulação de texto traduzido para ajustes e correções) e pesquisa (transformação de texto traduzido em dados numéricos). A cada uma delas corresponde a impressão de relatórios de situação e salvamento de arquivo.

O aplicativo converte tudo o que é possível e previsível no português brasileiro, para reduzir ao máximo o trabalho na edição. Fundamentalmente, o aplicativo trabalha com fonologia, mas há como o usuário editar o texto convertido para alguns aspectos fonéticos.

Para instalar o Nhenhém 1.0, basta copiar para o computador a pasta Nhenhém, dentro da qual estão o arquivo executável do programa, o arquivo com instruções iniciais leiame.txt e as subpastas Estatística e Fontes do IPA. O programa pode ser obtido por meio do sítio: <http://br.geocities.com/sisnhenhem>.

3.2 FINALIDADE

O programa de conversão grafema-fonema Nhenhém© versão 1.0 é um sistema eletrônico que converte a língua escrita (grafemas) em símbolos gráficos da língua falada (representação de fonema), a partir dos caracteres do International Phonetic Alphabet (IPA). O sistema trabalha com português do Brasil de acordo com a ortografia oficial vigente em 2008. Pode-se dizer que se trata de um sistema de conversão grafema-grafofonema, uma vez que ele não trabalha com sons, mas sim com representações gráficas deles.

3.3 REQUISITOS

O Nhenhém funciona em plataforma operacional Windows 98 e superiores, no entanto, as plataformas operacionais recomendadas são Windows 2000 SP4 e, principalmente, Windows XP. A fonte do IPA usada é a IPAPhon, que deve estar instalada no computador. Essa fonte deve ser copiada para a subpasta Fonts da pasta Windows. A configuração do teclado deve ser 'Teclado padrão com 101/102 teclas...' e o idioma deve ser português do Brasil nas configurações regionais do Painel de controle.

O usuário deve ter conhecimento em fonologia.

3.4 DESEMPENHO

Com o sistema, o usuário pode converter automaticamente desde uma palavra até um texto de 20 páginas, editá-lo, salvá-lo, pesquisá-lo e imprimi-lo. O sistema faz a conversão com acerto estimado de pelo menos 95%. Por isso, permite ao usuário editar o texto convertido para acertar os 5% ou menos de falha, trocar e inserir símbolos opcionais, ajustar dialetos etc.

O programa também permite gravar vários textos em um banco de dados para uso específico em relatório de estatística.

3.5 PRINCÍPIOS DE CONVERSÃO

Comandos padrão do sistema operacional Windows funcionam no Nhenhém. O texto é convertido à medida que o usuário o digita ou o cola. Os textos colados devem estar com formatação simples. A tonicidade é marcada mediante ordem do usuário. O sistema ignora tudo o que não é pronunciado (sinais de pontuação, símbolos, números) e converte todos os caracteres maiúsculos para minúsculos, por serem equivalentes na fala. Por esse motivo, a tecla CAPS LOCK não pode estar ativa durante a digitação ou colagem do texto. Algarismos indo-arábicos devem vir por extenso, para não ser

também ignorados. Isso vale para algarismos romanos, que não são ignorados, mas sim convertidos como se fossem letras em palavras. Assim ocorre com unidades de medida. Letras que não fazem parte do alfabeto do português podem gerar erro, pois o sistema tentará convertê-las ou as repetirá. Digitar espaços duplos entre palavras faz com que o cursor vá para o início do campo.

Combinações estranhas ao português do Brasil, como *chr*, não estão previstas, obviamente. No caso de *chr*, há conflito com as palavras que contêm "ch", como "chá".

O sistema trabalha com as letras típicas do alfabeto oficial vigente em 2008. Portanto, "w" e "y", em texto normal (antes de ser convertido), podem gerar erros de tradução, pois não há palavras no vocabulário português que contenham essas letras. Quando estrangeirismos que as contêm são aportuguesados, essas letras são substituídas neles. Para maior eficiência do sistema, no caso de nomes próprios, elas devem ser substituídas pelas letras correspondentes em português.

Também a letra "k" não faz parte do alfabeto oficial. A letra "k" é somente usada como consoante no português, mas seu uso também pode provocar outros erros. Novamente, aconselha-se usar a pronúncia do português (por exemplo: *skate* → "esqueite") ou eliminar do texto a palavra que gera problema. Em palavras, essas três letras figuram em casos especiais de nomes próprios de origem estrangeira, em derivações deles, em unidades de medida e em siglas. No entanto, a pronúncia é brasileira.

O sistema garante uma conversão possível de o que é previsível na língua, para o que há regras. Então, parte da língua não pode ser convertida propriamente. O que não é previsível – da mesma forma como ocorre com a revisão ortográfica e gramatical dos editores de texto eletrônicos – gera impropriedades que o usuário deve conhecer e retificar.

Desse modo, questões de harmonia vocálica, grafias cuja justificativa é puramente etimológica, ou seja, que não sofreram adaptação nas reformas ortográficas para ficar mais intuitivas e uniformes, não podem ser propriamente convertidas a todo tempo. Da mesma forma, certos encontros vocálicos em que a descodificação das letras

“e” e “o”, sobretudo em hiatos, prefira /i/ e /u/ devem ser editados no Nhenhém, caso não se queira adotar a pronúncia que o sistema apresenta. Quando há regras lógicas, o sistema faz a conversão automaticamente.

Na maioria das circunstâncias em que houver conflito, o sistema não converterá o grafema, a fim de reduzir erros de conversão. Nos casos em que a grafia da letra “x” não é previsível, o sistema não a converte, apenas a substitui. Cabe ao usuário editar o texto e manipular a conversão dessa letra. A maioria das letras “e” e “o” abertas sem acento gráfico permanecem com timbre fechado no texto convertido. Em alguns casos de ocorrência da letra “s” é necessário ajuste, pois há imprevisibilidade e incompatibilidade com outras regras (como no prefixo “trans” – e, por conseqüência, no radical “trans-” – e quando sucede “sub”).

Os controvertidos ditongos crescentes, à exceção de (q)“ua”, (q)“uo”, (g)“ua”, (g)“üe”, (g)“üi”, são tratados como hiatos. Isso significa que, no lugar da suposta semivogal /y/ ou /w/, na verdade, há uma vogal /i/ ou /u/. Esse princípio de funcionamento do Nhenhém é respaldado pela teoria lingüística (ver SAID ALI, 1964).

Nos ditongos, a semivogal /i/ é representada por /y/ em vez de /j/. Isso se deve ao fato de a letra “j” fazer parte do alfabeto original do português, e a letra *y*, não, de forma que usar “j” para semivogal provocaria erro em todas as palavras escritas com “j” e tornaria o sistema propenso a erros de descodificação. Se desejar usar /j/ em vez de /y/, o usuário pode editar o texto convertido. Depois de convertido, o texto não apresenta o símbolo /j/, então, não haverá mais problema.

A tonicidade marcada pela acentuação gráfica (acentos agudo e circunflexo), é responsabilidade do usuário. A posição da marca de tonicidade (/ˈ/) é antes da vogal tônica, o que a teoria mostra fazer sentido, no entanto, na edição, o usuário pode alterar a posição do símbolo de tonicidade. Apesar disso, aconselha-se a não interferir na posição da tonicidade, pois isso pode facilitar a ocorrência de erros de conversão, uma vez que os padrões silábicos são variáveis. O Nhenhém vale-se do fato de que, não importa o tipo de sílaba em questão, sempre há uma única vogal nela. Assim, ele consegue reproduzir o sistema de acentuação não marcada graficamente na íntegra. O

programa reconhece os principais clíticos da língua portuguesa, portanto, não marca sua tonicidade.

Ao digitar o texto, o usuário não deve digitar caracteres não legíveis. Se digitar algum, o cursor pulará para o canto superior esquerdo do campo ‘Digite ou cole...’. Assim, o hífen que une palavras com mais de um radical e afixos não deve ser digitado.

3.5.1 Símbolos Internalizados

Os símbolos internalizados no programa Nhenhém provêm da fonte IPAPhon.

Quadro 6: Letras e fonemas correspondentes no Nhenhém 1.0

Graf.	Fon.	Exemplo
á	/ˈa/	água
à	/ă/	àquela
â	/ˈã/	lâmpada
ã	/ã/	maçã
é	/ˈɛ/	pé
é	/ˈẽ/	também, contém
ê	/ˈe/	lêvedo
ê	/ˈẽ/	têmpora, ênfase
e	/ɛ/	era
e	/i/	cante
í	/ˈi/	lívida
í	/ˈ ĩ/	límpido, índio
i	/y/	peito
	/i/	ad(i)vento
ó	/ˈɔ/	pó
õ	/õ/	anões
ô	/ˈo/	pôs
ô	/ˈõ/	cômputo, cônscio
o	/ɔ/	somente
o	/o/	comente
o	/w/	mão
o	/u/	pato
u	/w/	pau, taquara
ú	/ˈu/	útil
ú	/ˈũ/	cúmplice, anúncio
ü	/w/	cinqüenta
c	/s/	acerola
c	/k/	acudir
ch	/ʃ/	achar
g	/ʒ/	gente, girar

Graf.	Fon.	Exemplo
gu	/g/	guerra, guitarra
h		hoje, ah
j	/ʒ/	janela
l	/w/	anzol
l	/l/	lenço
lh	/ʎ/	malha
lh	/l/	filhinho
m	/m/	mia
n	/n/	ano
nh	/ɲ/	ninho
qu	/k/	quente, quinta
q	/k/	aquático
r	/r/	cera, prata
r	|R|	amor
r	/ʀ̥/	bilro, enredo*
r	/ʀ̥/	rosto
rr	/ʀ̥/	amarrar
s	/s/	sapo
s	|S|	mosca, lesma
ss	/s/	assar
sc	/s/	fascinante
sç	/s/	nasça
s	/z/	asa
x	/kis/ ou /ks/	táxi
x	|S|	expor
x	/z/	exato
xc	/s/	exceção
z	/z/	azedo
z	|S|	luz

* Por conta de incompatibilidade entre a fonte do IPA utilizada e as configurações internas do computador, o símbolo correspondente ao fonema forte /ʀ/ adotado nessa versão do programa é /ʀ̥/.

3.5.2 Impropriedades

Em alguns computadores pode aparecer um espaço após as vogais nasais, que corresponde ao lugar ocupado pelo til, que sozinho é um caractere nas fontes do IPA. Esse espaço não existe – sua aparição depende das configurações do computador do usuário – e não aparece quando o texto for aberto no Word. Pode-se ignorá-lo e editar o texto normalmente.

Se o cursor permanece à esquerda da tela e o texto é digitado de trás para frente, desative a tecla CAPS LOCK. Se o texto colado não aparece no campo 'Resultado', verifique se não há um espaço em branco no final do texto colado no campo 'Cole ou digite...', apague esse espaço. Se mesmo assim o resultado não aparecer, selecione a opção 'Prosódia'.

Se aparecem caracteres impróprios em alguma palavra isolada no texto convertido, pode ser que antes dela, no texto original, haja uma combinação de ENTER e pontuação que o sistema não detectou. Procure essa palavra no campo 'Digite ou cole o texto a ser convertido' e apague os caracteres anteriores a ela, até que ela grude na palavra anterior, então, dê um espaço. A conversão ficará correta.

Se no texto convertido caracteres normais aparecerem com símbolos impróprios, pode ser que a configuração do teclado no Painel de controle do Windows esteja incorreta. Ainda pode haver incompatibilidade entre o código hexadecimal e o computador, que afeta alguns caracteres. Se muitos caracteres estranhos aparecerem na tradução, provavelmente, a fonte do IPA requerida não está instalada.

A presença de símbolos estranhos no texto convertido também pode significar que havia algum caractere ilegível e de uso raro no texto original, que o sistema não conseguiu ignorar. Algarismos romanos viram palavras estranhas na tradução. Recomenda-se retirar esses caracteres antes de converter o texto ou substituí-los por palavras.

Se o texto convertido salvo não abrir no Word, provavelmente, não foi colocada a extensão do arquivo na caixa de diálogo 'Salvar'. Salve-o novamente, digite o nome do arquivo, ponto e a extensão (exemplo: texto1.doc).

Se o tamanho da fonte não mudar quando for escolhido um tamanho diferente nos campos ‘Opção de fonte’, sobretudo na edição, o texto foi alterado antes dessa ação. Primeiramente, mude o tamanho da fonte, depois comece a editar o texto.

Cortes nas palavras no final da linha, nos relatórios de conversão, devem-se ao comportamento da fonte do IPA, que não funciona como as fontes normais de texto. Essa quebra de texto irregular também ocorre no texto convertido aberto no Word. A fonte IPAPhon não lê quebras de linha, por isso, trata o texto todo como se fosse uma única linha. O Nhenhém não interfere nisso. Edite o texto normalmente, depois, abra-o no Word para arrumar as separações incorretas e juntar as linhas, se for o caso. Na edição do Nhenhém, trocas de linha podem ser problemáticas. Esses cortes não interferem no relatório de estatística.

Se ocorrer um erro de exceção (mensagem em inglês) ao digitar ou colar um texto, pode ser que o texto colado anteriormente estava na fonte IPAPhon. A solução é fechar o Nhenhém para que a memória seja descarregada e reiniciá-lo em seguida. Para maior segurança do usuário, as fontes *true type* mais facilmente reconhecidas pelos computadores devem ser usadas no texto a ser colado. São elas: *arial*, *courier* e *times new roman*, a mais recomendável.

Em alguns casos, pode não aparecer o /i/ epentético entre consoantes.

Outros erros não documentados aqui devem ser informados à desenvolvedora do Nhenhém, para análise e correção.

3.5.3 Ajuste Obrigatório

Afora os casos em que é previsível, o grafema “x” é apenas substituído por /X/, para que o usuário o corrija. A pronúncia de /o/ e /e/ com timbre aberto (baixo) e sem acento gráfico também deve ser ajustada, pois o sistema converterá esses casos como timbre fechado (alto), na grande maioria das vezes. A ocorrência da letra “s” quando corresponder a /z/ após “trans”, “ob” e “sub” deve ser editada, pois o programa sempre a traduz como /s/. A transcrição da palavra “que” quando ocorre depois da preposição “por” deve ser ajustada também.

3.9 FUNCIONAMENTO

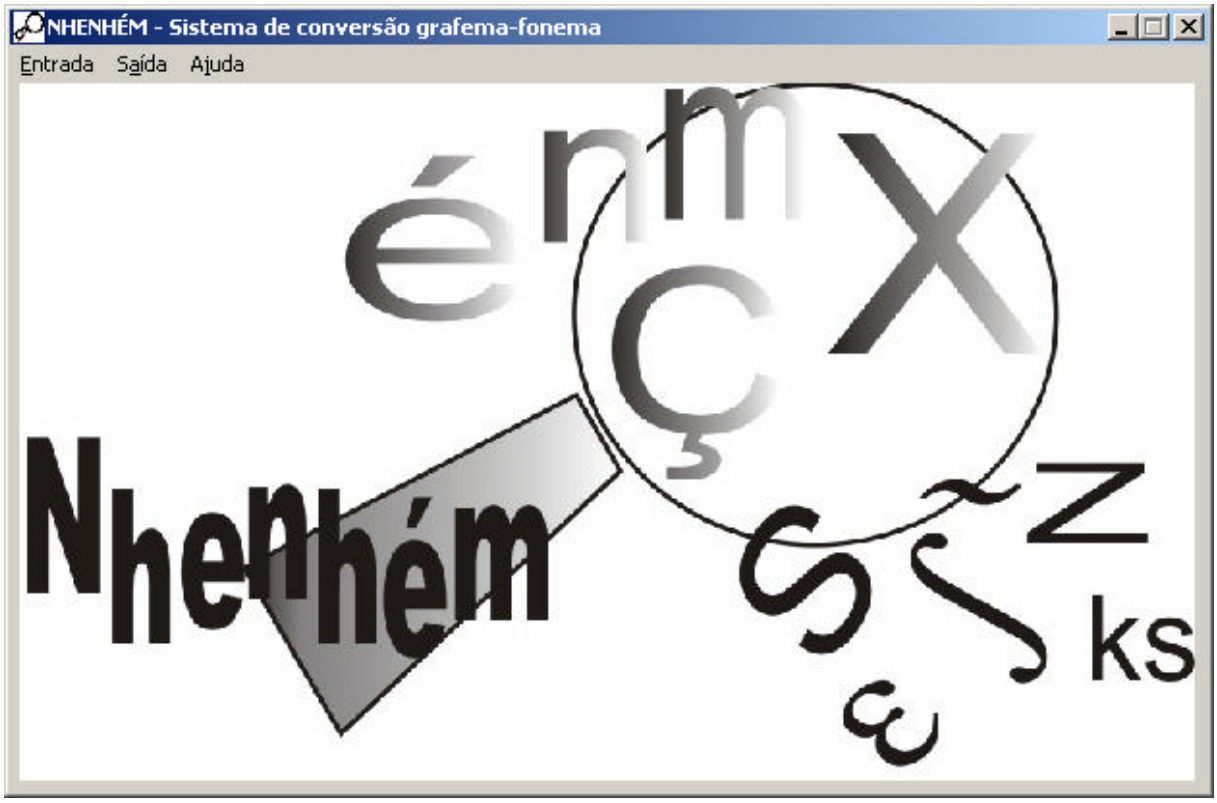


Figura 5: Tela principal do Nhenhém

Após desaparecer a tela de abertura, aparece a tela principal do sistema, da qual se acessam as demais opções de uso, a partir dos seguintes menus:

Entrada – Digitar ou colar: permite ao usuário trabalhar com conversão de texto.

Saída – Pesquisar: permite extrair estatística de um ou mais textos convertidos.

Ajuda – Sobre o Nhenhém: mostra informações sobre o sistema e créditos.

– Manual de instruções: abre o manual de instruções do programa.

– Contrato: mostra o contrato de uso do programa, a identificação do usuário e o número da licença.

Os textos a ser convertidos são as entradas do sistema, os textos já convertidos são as saídas. Os comandos dos menus também podem ser acessados pelas teclas de atalho F2 e F3. Para sair do sistema, pode-se clicar no ícone do programa, na barra de título da tela, no canto superior esquerdo, e selecionar a opção Fechar, ou clicar duas vezes nesse ícone, ou ainda clicar no botão à direita da barra de título.

3.9.1 Conversão

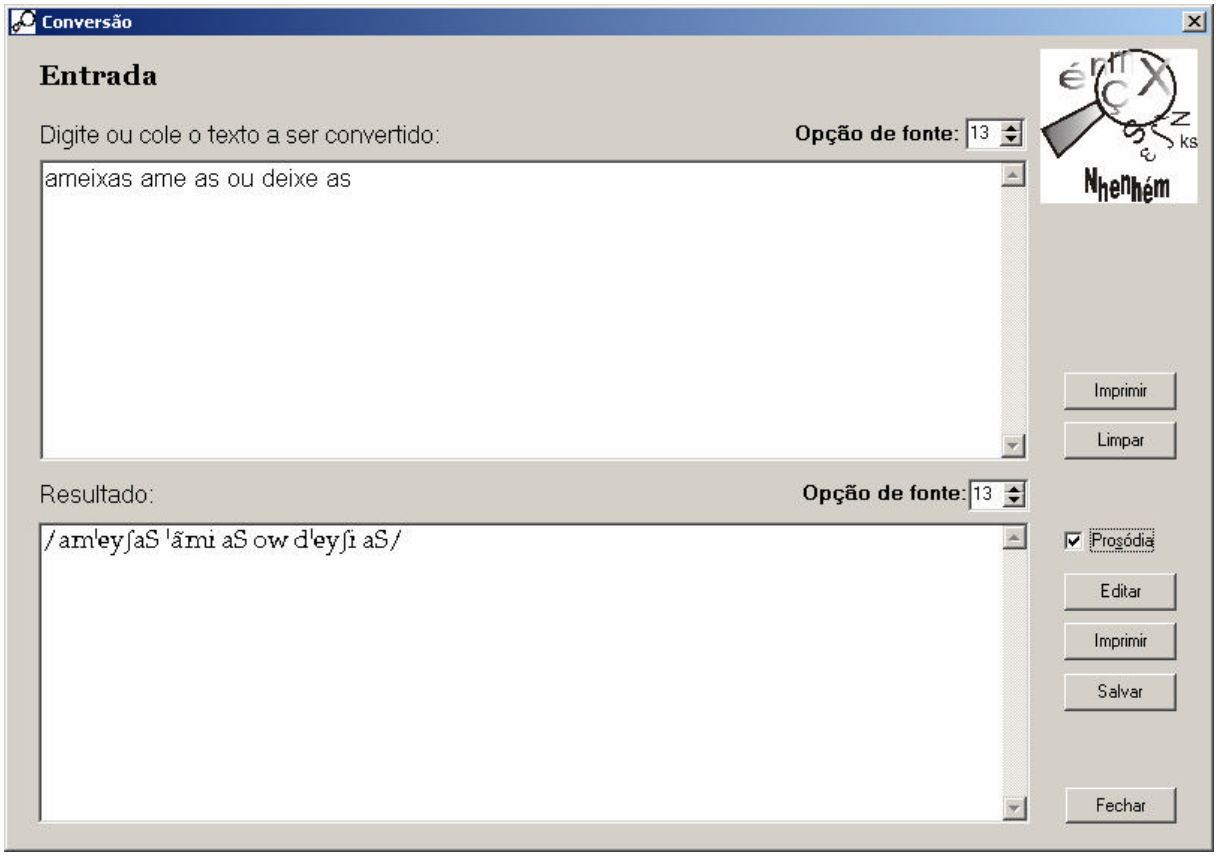


Figura 6: Tela de conversão do Nhenhém

Na tela **Entrada** o usuário insere o texto que deseja converter no campo ‘Digite ou cole o texto a ser convertido’. Simultaneamente à escrita, o texto aparecerá convertido no campo ‘Resultado’. Em caso de texto colado e extenso, o resultado pode demorar alguns segundos a aparecer. Os campos ‘Opção de texto’ servem para aumentar ou diminuir a fonte em que os textos original e convertido aparecem nas

janelas. Textos colados não devem ter formatação complexa, porque ela pode inibir os botões da tela **Entrada**.

Os botões à direita da tela têm a seguinte função, respectivamente:

- **Imprimir –** imprime o texto a ser convertido, para o usuário ver os sinais que o sistema ignora na conversão (como maiúsculas, pontuação, parênteses, números), e compará-lo com o texto convertido. Esse relatório serve para guiar o usuário na checagem do texto convertido.
- **Limpar –** limpa o campo 'Digite ou cole...', para nova digitação ou colagem.
- **Prosódia –** mostra a tonicidade no texto do campo 'Resultado'.
- **Editar –** abre a tela **Edição de texto convertido**, já com o texto convertido à mostra, para ser editado.
- **Imprimir –** mostra a visualização de impressão do texto convertido em relatório preliminar.
- **Salvar –** permite salvar o texto convertido em disco, em formato de arquivo do editor de textos Word. A extensão do arquivo (.doc) deve ser digitada pelo usuário nas operações de salvamento.
- **Fechar –** fecha a tela **Entrada** e volta para a tela principal.

3.9.2 Edição

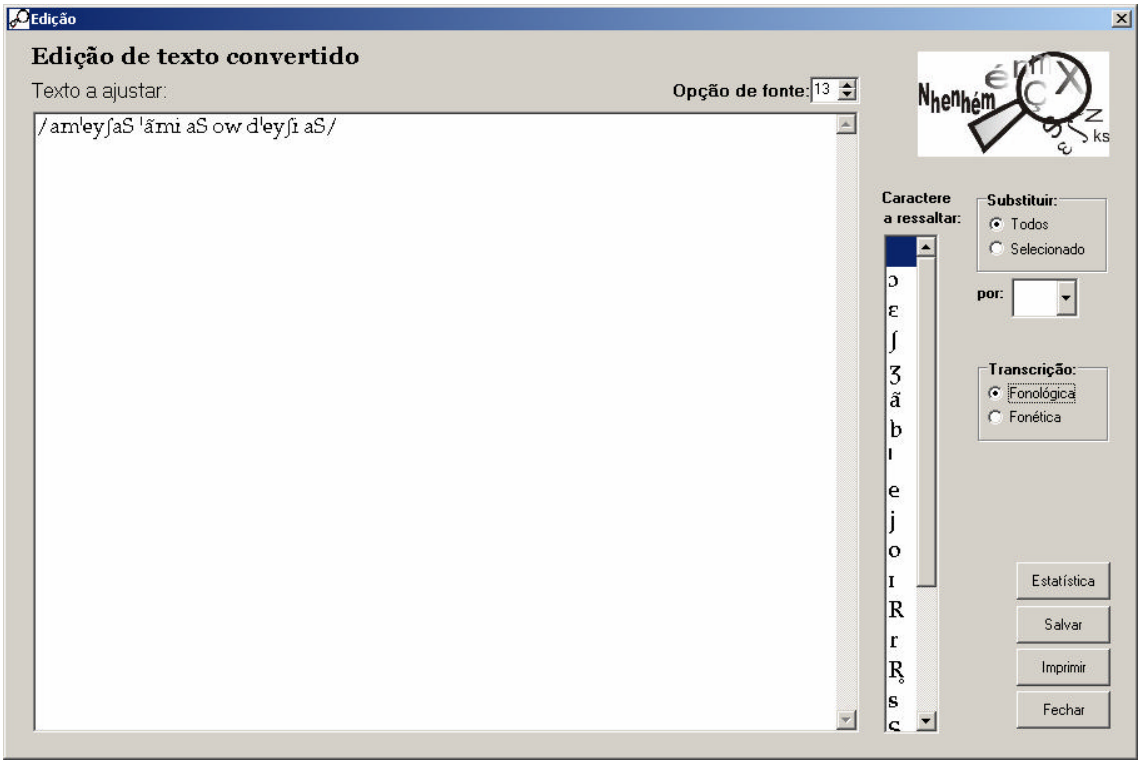


Figura 7: Tela de edição do Nhenhém

A formatação do texto, como eliminar linhas em branco, se houver, agrupar sentenças em uma linha, juntar palavras eventualmente quebradas na conversão, é trabalho do usuário. A formatação somente deve ser corrigida se isso for estritamente necessário. Para fins de montar um banco de dados, ou seja, um *corpus* fonológico, aconselha-se a não formatar o texto, apenas ajustar impropriedades de conversão. Se houver necessidade de formatar o texto, deve-se salvá-lo em formato para ser aberto pelo editor de textos Word. Em casos em que as duas operações forem necessárias, salva-se uma cópia do texto para estatística (que não pode ser ajustada em sua formatação) e outra para ser formatada nos editores de texto do Windows.

A tela **Edição** serve para editar o texto convertido, portanto, digitar palavras ou colar caracteres e texto diretamente na **Edição** não implica conversão deles. Após salvo, o texto convertido também pode ser ajustado no editor de textos Word, se o usuário usar o método de inserir símbolos desse editor. O arquivo gerado nesse caso não servirá para análise estatística. Mais importante da **Edição** é ajustar caracteres

fonéticos. Pode-se ajustar um único caractere por vez ou todos os do mesmo tipo ao mesmo tempo.

Para ajustar todos, selecione o caractere no campo 'Caractere a ressaltar'. Todos os caracteres correspondentes no texto ficarão destacados em vermelho. No campo 'Substituir', a opção 'Todos' deve estar marcada. Então, no campo 'Por', escolha o caractere que ocupará o lugar dos que estão em vermelho. Para mudar os caracteres um a um, selecione o caractere no texto com o *mouse* ou pelo teclado. No campo 'Substituir', a opção 'Selecionado' deve estar marcada. Então, no campo 'Por', escolha o caractere que ocupará o lugar daquele que estava marcado no texto.

Para inserir um caractere do IPA no texto convertido, como o /i/ epentético, basta posicionar o cursor no local onde entrará o símbolo desejado, no grupo de opção 'Substituir', marcar 'Selecionado', e escolher o caractere correspondente na caixa de combinação 'Por'.

Em caso de o usuário fazer a edição e usar caracteres que deixam o texto fonético em vez de fonológico (como ajustar o texto para corresponder a um dialeto específico), ele deve escolher a opção Fonética, no grupo de opção 'Transcrição'. Nesse grupo, a opção padrão é 'Fonológico', conforme a proposta inicial do Nhenhém.

Na tela **Edição de texto convertido**:

- **Estatística –** permite salvar o texto editado em formato especial para que a ele seja aplicada estatística futuramente.
- **Salvar –** permite salvar o texto editado em disco, em formato de arquivo do editor de textos Word. A extensão do arquivo (.doc) deve ser digitada pelo usuário nas operações de salvamento.
- **Imprimir –** permite imprimir o texto editado, para o usuário checar os ajustes feitos.
- **Fechar –** permite sair da tela **Edição**.

Ao salvar o texto para estatística, o sistema mostra a seguinte mensagem de confirmação: "Arquivo NH...vvo salvo", se o texto for fonológico, ou "Arquivo NH...vve salvo", se o texto for fonético. Isso significa que o texto foi para o banco de dados, na pasta Estatística.

Leva certo tempo para se acostumar com as operações de edição, sobretudo de ajuste, mas é questão de prática e disciplina. É mais fácil lidar com algumas ações no Word, mas, mesmo nesse caso, primeiramente, aconselha-se a fazer as operações de ajuste de texto pós-conversão no Nhenhém.

Quanto aos textos para estatística, o programa nomeia cada arquivo com as letras "NH" seguidas pelos dados do momento em que foi feita a gravação e salva os arquivos com extensão .vvo ou .vve, que designa verificação estatística fonológica (opção padrão) ou fonética (por escolha do usuário). Por exemplo, se um arquivo foi criado no dia 16 de junho de 2008, às 15h18min25s, o sistema grava-o com o nome: NH160608151825.vvo, para que não haja risco de haver dois arquivos gravados no mesmo minuto, pois isso implicaria a exclusão do arquivo existente.

Depois de salvar os textos para estatística, não se sabe mais qual é o conteúdo de cada um deles, portanto, sugere-se que o usuário mantenha controle paralelo dos conteúdos, dependendo da relevância de cada texto. Ele pode fazer uma listagem em que constem o nome do arquivo e uma dica que lembre qual é o conteúdo ou gravar os arquivos de origem com o mesmo nome dado pelo Nhenhém na transcrição correspondente, por exemplo. O programa trabalha fundamentalmente com fonologia, portanto, certos ajustes fonéticos podem interferir no relatório de estatística. Não há problemas com ajustes de vogais. No caso das consoantes, pode haver problemas maiores. Ao fazer ajuste fonético, o usuário deve usar somente os símbolos que o relatório de estatística lê, se quiser aplicar análise numérica ao texto ajustado.

O usuário deve testar o programa, adaptar-se a seu funcionamento, para então montar o banco de textos.

3.9.3 Pesquisa

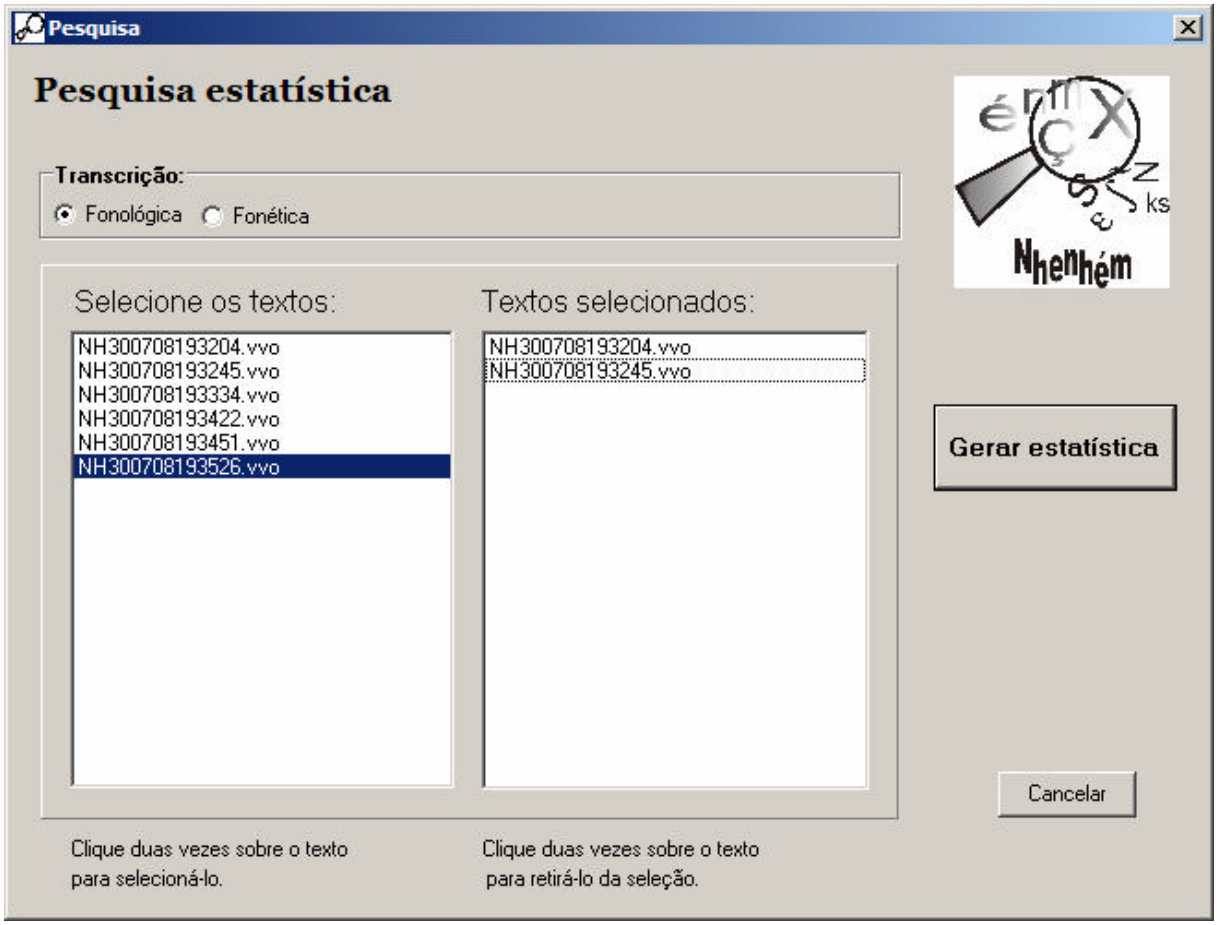


Figura 8: Tela de pesquisa do Nhenhém

A pesquisa serve para o usuário conhecer as características fonológicas ou algumas características fonéticas de um texto ou de vários textos, em números inteiros e em porcentuais. No menu 'Saída' da tela principal há a opção 'Pesquisar', que abre a tela **Pesquisa estatística**. Nessa tela, a primeira opção a escolher refere-se ao tipo de transcrição dos textos a analisar: fonológica (que mostra os arquivos com extensão .vvo) ou fonética (que mostra os arquivos com extensão .vve).

Na caixa de listagem 'Selecione os textos' aparecem os textos que compõem o banco de dados do usuário, aos quais será aplicada a avaliação estatística. A caixa de listagem 'Textos selecionados' mostra os textos que o usuário selecionou para o programa analisar numericamente. O usuário controla os textos a ser analisados ou não mediante dois cliques no campo desejado. A forma indicada para eliminar um arquivo

do banco é selecioná-lo e pressionar a tecla BACK SPACE. Os arquivos são excluídos um a um, e não podem ser recuperados.

Ao se pressionar o botão **Gerar estatística**, o sistema analisará os caracteres dos textos em virtude de seus traços distintivos, convertê-los-á em números inteiros e porcentuais e mostrará o resultado em um relatório, que pode ser impresso.

O limite de textos do banco de dados não foi testado. O usuário será informado pelo sistema quando não for mais possível carregar textos para gerar estatística.

3.9.3.1 Relatório estatístico fonológico ou fonético

A partir da tela **Edição**, o programa grava os arquivos para análise numérica dentro da subpasta Estatística. Essa é a pasta da massa de dados fonológicos. Esses arquivos não devem ser abertos pelo usuário, porque contêm formatos internos específicos de trabalho do programa. Editá-los pode prejudicar a confiabilidade da pesquisa estatística.

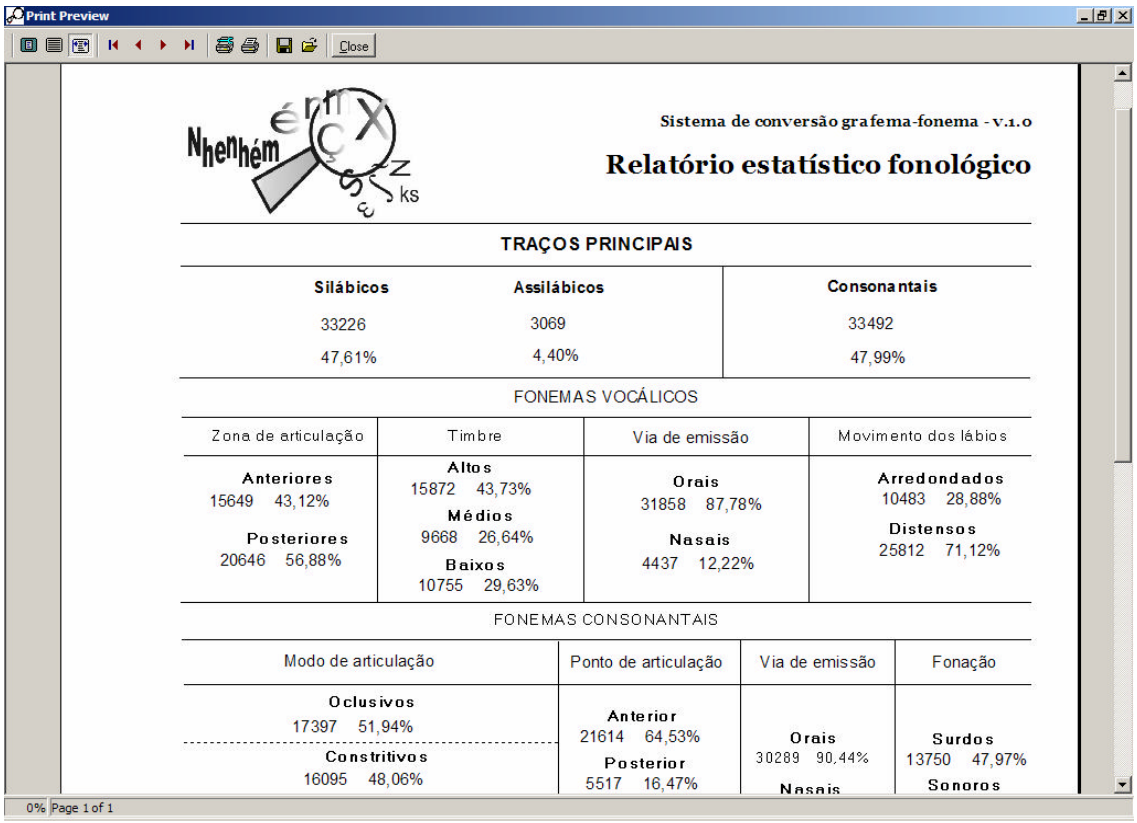


Figura 9: Visualização de relatório de estatística do Nhenhém

O relatório estatístico fonológico e fonético fornece dados numéricos acerca dos fonemas. São fornecidos a quantidade encontrada em números inteiros e esse valor transformado em porcentual.

O relatório de estatística desconsidera a tonicidade das palavras e o símbolo /X/, que não deve constar nos textos do banco. Ao texto ser inserido no banco, a tonicidade das palavras que ele contém já fez seu papel, por isso e por não ser legível, ela não interfere na contagem dos fonemas. Entretanto, há fenômenos fonológicos que dependem da prosódia, de forma que não se deve salvar para estatística um texto sem marcas de tonicidade.

O relatório divide-se em quatro partes: Traços principais; fonemas vocálicos; fonemas consonantais; e porcentuais e quantias individuais de fonemas. Os números correspondentes a essas seções, ao ser gerado o relatório, originam-se dos fonemas constantes nos textos selecionados na tela **Pesquisa estatística**.

Traços principais

Os traços principais dos fonemas são os silábicos (propriedade das vogais, que podem ser ápice de sílaba), assilábicos (propriedade das semivogais, que não podem ser ápice de sílaba, nem iniciam sílabas, mas se associam a uma vogal na mesma sílaba) e consonantais (propriedade exclusiva das consoantes, que se associam a uma vogal na sílaba).

Cada sílaba pode ter mais de uma consoante (“creme” → /krˈemi/), até duas semivogais (“quais” → /kwayS/), mas somente uma vogal. As semivogais pertencem ao grupo das vogais, por seus traços fonológicos assemelharem-se aos das vogais /i/ e /u/.

Fonemas vocálicos

A divisão Fonemas vocálicos fornece quantias e estatística referentes às vogais e semivogais nas seguintes classificações:

- **Zona de articulação** refere-se à posição da língua na dimensão horizontal.
- **Timbre** refere-se à altura da língua (posição vertical) durante a pronúncia da vogal.
- **Via de emissão** refere-se ao caminho percorrido pela corrente de ar durante a vocalização.
- **Movimento dos lábios** distingue as vogais para cuja pronúncia os lábios contraem-se em sentido de fechar-se na altura central (arredondadas) das outras.

Fonemas consonantais

A divisão Fonemas consonantais trata das consoantes, segundo as seguintes classificações:

- **Modo de articulação** refere-se à maneira pela qual se forma o obstáculo à corrente de ar que dá origem ao fonema.
- **Ponto de articulação** é o lugar onde se forma esse obstáculo.
- **Via de emissão** refere-se ao caminho percorrido pela corrente de ar durante a vocalização. Pode ser simples (oral) ou dupla (oral e nasal), que é considerada nasal.
- **Fonação** refere-se à vibração ou não das cordas vocais durante a produção do som. Os arquifonemas |S| e |R| não estão incluídos nesses números, porque neutralizam esse traço.

Porcentuais e quantias individuais de fonemas

Essa divisão mostra cada fonema encontrado nos textos analisados, sua respectiva porcentagem e quantidade de aparições. Nos porcentuais individuais, pode haver diferenças de até 1% no total, dadas a arredondamento interno do computador, uma vez que não se usam casas decimais, por questão de espaço. Por isso, ideal é trabalhar com grande massa de fonemas, para que essa diferença se torne cada vez menos importante, bem como para que testes de distribuição sejam mais confiáveis. Os testes estatísticos podem ser validados com sua aplicação a outras massas fonológicas.

Se o resultado da estatística for diferente de o que se espera, deve-se verificar se os textos que a originaram foram editados corretamente antes de ser gravados no banco de textos. Se os textos estiverem corretos, pode-se estar diante de um fato lingüístico que merece investigação apurada.

A eficiência dos dados obtidos na estatística depende da formação adequada da massa fonológica, de acordo com a teoria geral sobre lingüística de *corpus* e sobre *corpus* lingüístico fonológico. Se os pressupostos dessas teorias não forem levados em conta, os resultados da pesquisa serão mais propensos a erros, mais criticáveis e, sobretudo, menos científicos.

3.9.4 Impressão

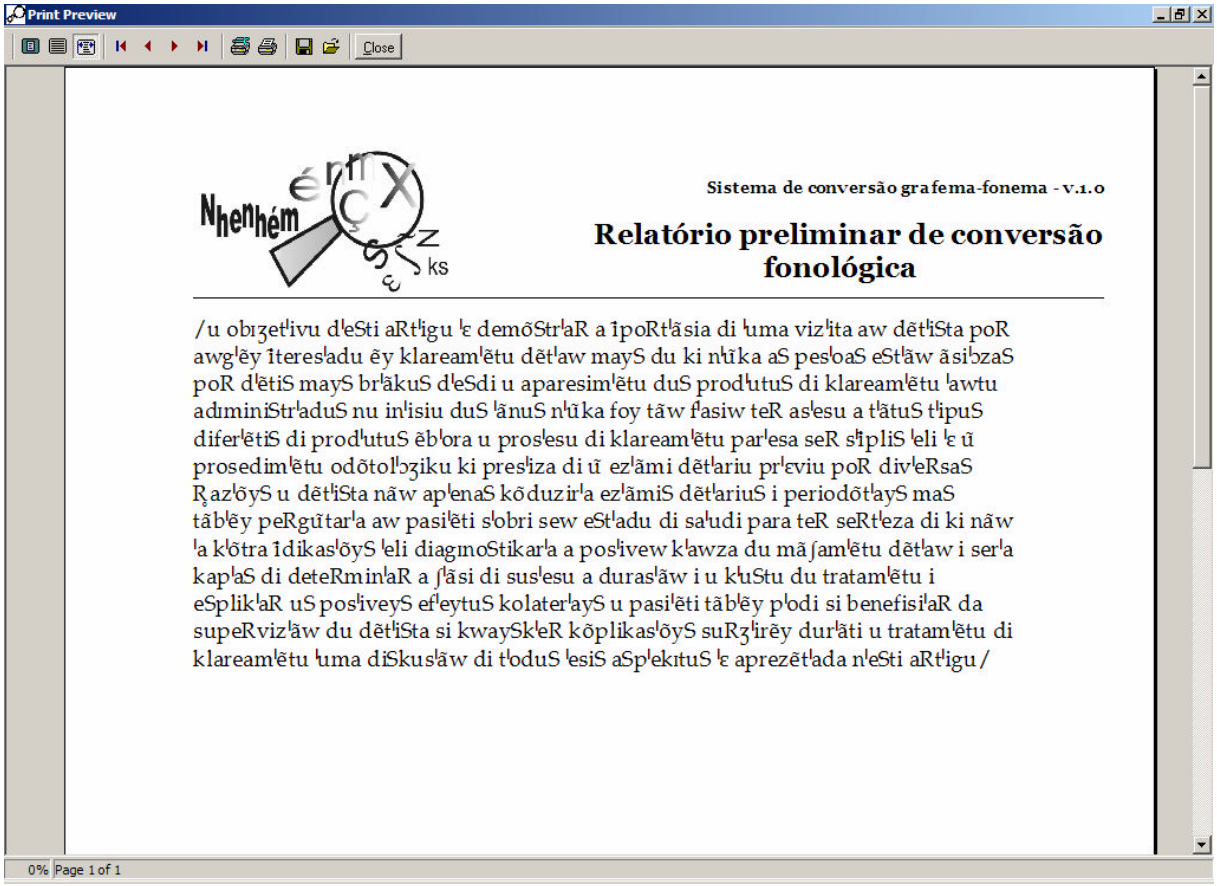


Figura 10: Relatório preliminar de conversão do Nhenhém

A impressão de relatórios é acessada pelo botão **Imprimir**, na tela que está aberta e corresponde à impressão de o que a tela permite. Os relatórios aparecem em modo visualizar impressão e podem ser impressos ao pressionar-se o botão **Print**, que traz a figura de uma impressora, na tela **Print view**. Os relatórios do sistema Nhenhém são: Relatório de texto para conversão, Relatório preliminar de texto convertido (acessados na tela **Entrada**); Relatório de ajuste de texto convertido (acessado na tela **Edição**); e Relatório estatístico fonológico ou fonético (acessado na tela **Pesquisa**).

4 O NHENHÉM EM DISCUSSÃO

ERRA UMA VEZ

nunca cometo o mesmo erro
duas vezes
já cometo duas três
quatro cinco seis
até esse erro aprender
que só o erro tem vez

Leminski (2008)

Uma vez que já se apresentou o Nhenhém, é tempo de expor implicações técnico-científicas advindas de seu desenvolvimento e aplicações do programa.

4.1 APRENDIZADO DA PRÁTICA

As convenções lingüísticas adotadas no Nhenhém foram, dentre as aceitáveis por alguma teoria, as mais lógicas, segundo mostraram testes feitos antes de iniciar a programação. Esse juízo foi fundamental para o programa cumprir sua finalidade e a pesquisa atingir seu objetivo. O computador, dada sua natureza, mostra o que é lógico e o que não é no sistema escrito em relação ao sistema falado, em vários aspectos. Desse distinto ponto de vista, emerge matéria para reflexão.

Na teoria, muitas das tabelas de fonemas existentes apresentam características não raro ilegíveis ou desnecessariamente complexas para ser inseridas em computador. Por exemplo, em um programa eletrônico, é incomum lidar com um fonema que apresenta os traços +anterior e -posterior, porque isso não é logicamente conversível em 0 e 1. Essa notação em + e - antes da característica fonológica não é apropriada para a programação de computador. O computador aceita os fonemas que são posteriores ou anteriores, um em oposição ao outro ou em oposição a uma terceira classificação. Por exemplo, os fonemas /p/ e /b/, em virtude de seus traços fonológicos, recebem a seguinte notação originada nas tabelas internas do Nhenhém:

/p/ 10101001000000000

/b/ 10011001000000000

O terceiro e o quarto caracteres das cadeias binárias acima correspondem aos traços surdo e sonoro, respectivamente, uma vez que eles distinguem os dois fonemas, e são a única diferença entre as duas cadeias. Ambos os fonemas contêm os dois traços, que se alternam em verdadeiro (1) e falso (0).

Notações complexas provocam erros e desvios nos cálculos e dificultam a interpretação deles. Há como desenvolver uma lógica para inseri-las em um programa, mas essa lógica é desnecessariamente trabalhosa e pode se tornar frágil, portanto, passível de provocar inconsistências futuras e dificultar adaptações e correções a ser feitas no programa. O melhor caminho é o mais simples e seguro. Nesse sentido, uma tabela fonética ou fonológica deve conter traços pertencentes e não pertencentes ao fonema. À medida que se desenvolvem os cálculos, afunilam-se as possibilidades e o próprio aplicativo mostra como as regras e os traços devem ser inseridos no ambiente eletrônico.

As situações devem ser simples o suficiente para ser lógicas. Esta experiência mostra que a falta de lógica – ou coerência – em muitas notações e definições prejudica o ensino, pois os professores não conseguem entender o sentido das classificações, conseqüentemente, repetem teorias, sem explicações profundas, e abandonam a fonética e a fonologia no ensino de língua materna.

Várias tabelas diferentes de traços fonéticos e fonológicos foram mostradas a professores de alfabetização, a fim de verificar a compreensão que tinham delas. Eles deveriam escolher uma. Todos se identificaram com as mais objetivas. Da mesma forma, a classificação das semivogais (como consoante ou como vogal) levou em conta a intuição de professores, já que teorias aceitam as duas formas ou não tomam posição. Os professores foram unânimes em identificar os traços das vogais nas semivogais. Adotou-se essa intuição como parte da teoria usada no programa.

Quando se trata de reduzir tudo a 1 e 0, a fonética se mostra mais maleável do que a fonologia, na qual falta certo nível de objetividade nas abstrações. Dentre as áreas com as quais este estudo lida, a estatística é a que mais exige delimitações, ou seja, requer que se assuma um ponto de vista específico, e com a qual, conseqüentemente, o computador se mostra mais afim. A estatística demanda objetividade, afinal, é matemática. É importante analisar corretamente a estatística gerada, de modo a extrair validade científica dos dados apresentados.

Exemplo de não-correspondência exata entre fonologia e matemática é o uso de arquifonemas. Isso ocorre por causa de traços distintivos faltantes aos arquifonemas, como surdo e sonoro, porque, na estatística, obrigatoriamente deve haver um todo, os 100%. Na fala, esse traço existe ou não, o falante vai fazer uma opção sempre. Dessa maneira, o uso de arquifonemas pode disfarçar os padrões de distribuição fonêmica da língua analisada, o que prejudica resultados estatísticos.

Não há arquifonemas na língua oral, no entanto, no ensino, é capital a importância de abranger variedades. Também, é mais fácil criar um programa eletrônico que faça conversão fonológica do que um que faça conversão fonética, automaticamente. No entanto, para maior eficácia do programa Nhenhém, novos estudos devem ser desenvolvidos, no intento de reduzir e esmerar o uso de arquifonemas.

O fato de não haver palavras em português padrão escritas com as letras “w” e “y” facilitou a sistematização do programa no que tange às semivogais, o que eliminou dias de pesquisa e testes, que foram usados para resolver outras questões. A tentativa prévia de utilizar o símbolo “j” em vez de “y” mostrou que tal uso gera inconsistências no programa e dificulta sobremaneira o desenvolvimento de regras simples. Na computação, também vale a assertiva de que, se há dois caminhos igualmente válidos, o mais fácil deve ser tomado, porque evitará conflitos futuros.

Merece destaque a forma como foi desenvolvida a aplicação de análise estatística no Nhenhém, pois o programa simula internamente um sistema gerenciador de banco de dados. Por isso, não é necessário instalar um gerenciador (como o Access© e o Paradox©), porque o programa é preparado para ler dados externos

específicos, que não fazem parte dele, e compará-los com suas tabelas internas. Assim, se quaisquer arquivos do banco de dados forem apagados, o Nhenhém continuará funcionando perfeitamente, e o banco poderá ser refeito pelo usuário.

O maior desafio dentro de o que estava previsto no projeto da tese foi lidar com vogais, sobretudo com os encontros vocálicos (ditongos e hiatos). Esse obstáculo foi transposto com adoção de posturas o mais lógicas, dentre as opções teóricas disponíveis, adaptações e tomadas de decisão. Fundamental foi tomar decisões extremas, do tipo sim ou não, porque isso possibilitou criar regras. Ainda assim, considera-se que casos de redução de "e" e "o" para /i/ e /u/ em posição pretônica podem ser mais bem resolvidos no Nhenhém, em versões futuras.

Nesse sentido, encontros vocálicos cuja pronúncia é flutuante merecem análise mais profunda antes de se criar uma sistematização diferenciada para eles, ou há risco de ela não corresponder a alguma das realidades de pronúncia. É preciso ter cautela e sempre considerar o conjunto dos tipos de ocorrências, e não exemplares aleatórios delas.

Por exemplo, o encontro vocálico escrito "ae", que ocorre em "caetano", "paetê" e "baeta", pode ser lido como ditongo no primeiro caso: [kaytˈãnu] (mas também como hiato: [kaitˈãnu]), parece hiato no segundo: [paetˈe] ou [paitˈe], mas é claramente hiato no terceiro caso: /baˈeta/. Mais comum em português do Brasil é haver hiato nessas circunstâncias, ou seja, segundo e terceiro casos. Em situação similar, quando o ditongo é claro, o sistema escrito prefere a grafia "ai" a "ae", que é rara. Tanto é assim que o dicionário Aurélio (1999), que contém 435 mil verbetes, registra 130 verbetes em que há o encontro vocálico "ae" e 2.589 verbetes em que há o encontro "ai". São escassos os casos em que "ai" pode não ser descodificado como ditongo, normalmente, são vocábulos em que há composição, como "reidratado". Mesmo assim, a pronúncia é flutuante. Esse caso refere-se ao encontro vocálico interno à palavra, quando está no final dela, a descodificação é mais uniforme.

Isso demonstra que estudos aprofundados e decisivos (em que se assuma uma posição unilateral) sobre ditongos no português do Brasil devem ser feitos. O Nhenhém adota uma postura fixa em relação a todos os encontros vocálicos, mas

alguns deles configuram casos flutuantes, que merecem apuro teórico, para que a prática (a conversão do programa) não precise ser editada.

Fazer análises quantitativas e qualitativas lexicais e textuais sobre encontros vocálicos seria um ponto de partida melhor do que analisar casos isolados, pois este método ignora outros casos que, quando ocorrem, fragilizam as teorias. Por meio das análises quantitativas e qualitativas, podem-se conhecer todos os tipos de encontros vocálicos que ocorrem em português, as circunstâncias em que ocorrem, as flutuações de pronúncia e freqüência de cada um na língua. As propostas advindas de tais análises poderiam nortear as convenções escritas, que devem ser lógicas e intuitivas ao usuário da língua, tanto quanto for possível. Estima-se que o Nhenhém auxilie em estudos específicos sobre encontros vocálicos. No programa, o usuário pode criar *corpus* específico para trabalhar com essa questão.

Sistematizar a tonicidade da língua portuguesa do Brasil no Nhenhém foi um dos maiores desafios desta pesquisa, em termos de fonologia, porque isso não era previsto no projeto da tese. No entanto, nesse caso, o êxito foi de 99,90%, apenas com base nas terminações das palavras, que regem os vocábulos não acentuados graficamente, já que os acentos gráficos são inseridos pelo usuário.

A prosódia ensina. Ao ser estruturada no aplicativo, a prosódia ampliou a possibilidade de sistematizar as regras de descodificação do sistema ortográfico do português do Brasil. Somente então se revelou, por exemplo, que as terminações "el" e "ol" descodificam-se como /ɛw/ e /ɔw/ em palavras oxítonas, ou seja, quando forem tônicas. Assim se tem "sofrível" → /sofrˈivew/ e "carretel" → /kaɾetˈɛw/, bem como "boldo" → /bˈowdu/ e "caracol" → /karakˈɔw/. No último caso, a exceção fica por conta de "gol", que é aportuguesamento anômalo da palavra inglesa *goal*, entretanto, por ter apenas uma vogal, essa exceção não gera erro no programa. A partir desse achado, outros casos de timbre aberto não marcado graficamente podem ser resolvidos, bem como a redução de "e" e "o" para /i/ e /u/ em encontros vocálicos.

Também a ortoépia revelou que a letra "x", em sílaba final de palavra oxítona terminada em "i" equivale ao fonema /ʃ/ na leitura. Essa regra abrange palavras como "abacaxi", "xixi", "tucuxi", cujas traduções devem ser editadas por enquanto:

/abakaXˈi /, /ʃiXˈi/ e /tukuXˈi/. Essa e muitas outras regras podem ser elaboradas com apoio da prosódia, porque ela influencia os valores que os grafemas têm na descodificação, assim, contribui com a transparência do sistema alfabético do português.

Quanto à lista de exceções do programa, ideal é que seja a menor possível. Quanto maior ela for, maior será a limitação do programa, pois um sistema eletrônico deve trabalhar com regras, para dar conta de analisar qualquer palavra, e não procurá-las em uma biblioteca. As palavras que estão na lista não são analisáveis, são tabeladas, e isso implica que derivações delas não são corretamente traduzidas pelo sistema, embora representem as mesmas circunstâncias.

Algumas decisões tomadas na feitura do sistema são criticáveis a uns e louváveis a outros, indubitavelmente, bem como alguns teóricos escolhidos. No entanto, isso não foi optativo. As divergências são naturais, quando se unem áreas distintas. As escolhas vieram da necessidade imposta pela programação e, dentro disso, da objetividade e inteligibilidade das teorias existentes e da opinião e intuição de professores, alunos e outros usuários da língua. A eficiência do Nhenhém corrobora a utilidade das teorias adotadas.

4.2 FUNCIONALIDADE

Ao aplicar análise estatística no banco de dados fonológico montado com seis textos da área odontológica, obtiveram-se algumas informações relevantes acerca da língua. O programa analisou os textos em grupo e individualmente, o que gerou sete relatórios estatísticos. Os relatórios individuais confirmam o relatório unificado e vice-versa, pois os valores referentes à distribuição dos fonemas aproximam-se em todos os relatórios.

Disso se tem que os fonemas silábicos aparecem na leitura na ordem de 47%, os consonantais, 48% e os assilábicos (semivocálicos) ficam em torno de 4%, apenas com variações de casas decimais, que individualmente não chegam a 1% (Anexo 3). Isso confirma, estatisticamente, a uniformidade na distribuição fonêmica e que CV é a

sílaba mais comum do português. A quantidade de semivogais equivale à quantidade de ditongos que há no texto, pois elas somente figuram em ditongos.

Quanto aos fonemas vocálicos, destaca-se que os posteriores ocorrem em torno de 15% a mais do que os anteriores. Os posteriores que mais aparecem são /a/ e /u/, e dentre os anteriores é o /i/, que ocorre apenas 2% a menos do que /a/, que é o que mais ocorre dentre os fonemas vocálicos.

Quanto aos fonemas consonantais, há equilíbrio na ocorrência de constritivos e oclusivos, embora os oclusivos ocorram sempre em torno de 3% a mais do que os constritivos.

A aproximação dos resultados dos relatórios numéricos confirma, de certa forma, a confiabilidade da estatística do Nhenhém. A verificação dos relatórios de conversão correspondentes a cada texto da massa lingüística odontológica confirma que a tradução foi feita corretamente, porque os desvios de tradução encontrados foram previstos. Disso se tem que o procedimento para montagem de *corpus* fonológico descrito em seção anterior desponta como eficiente e que, numa massa fonológica, o tipo de textos analisados tende a ser mais importante do que a quantidade deles, respeitadas quantidades mínimas, salienta-se.

Uma vez que se sabe o que esperar nos resultados, ou seja, conhecem-se os padrões de distribuição da língua, a atenção pode concentrar-se em desvios e em diferenças nas casas decimais da estatística. Textos de crianças em fase de alfabetização podem desviar-se desses números, e cabe descobrir onde estão os desvios, para então se desvendar sua motivação. A análise aqui apresentada foi feita com uso de um *corpus* de textos técnico-científicos de uma área específica. Eles contêm palavras técnicas, portanto, muitas delas podem ser não usuais. Textos de outros tipos podem causar variações leves nos percentuais encontrados e mostrar erros de tradução que ainda não apareceram nos testes. Resta conferir. A visão e o preparo do usuário do programa enriquecem os números, criam formas diferentes de olhar esses dados e indicam soluções para problemas.

O relatório quantitativo apresenta outros dados dos quais se podem extrair muitas outras informações. No entanto, eles não foram mencionados aqui, porque o

objetivo era validar o *corpus* montado e a confiabilidade do relatório estatístico fonológico. A validação se efetivou.

A partir do relatório preliminar de conversão de um dos textos da massa, analisaram-se os casos em que /e/ e /o/ passam a /ɛ/ e /ɔ/. Nesse relatório, 1% correspondia a aproximadamente 130 ocorrências. A situação inicial e pós-ajuste são:

Tabela 3: Valores preliminares dos fonemas /ɛ/ e /ɔ/

Situação inicial						
e	6%	826		ɛ	1%	73
o	3%	397		ɔ	0%	45

Tabela 4: Valores dos fonemas /ɛ/ e /ɔ/ após ajuste

Situação final						
e	6%	759		ɛ	1%	140
o	3%	352		ɔ	0%	84

Os resultados finais não modificam os valores estatísticos, considerando-se o arredondamento de casas decimais. Isso indica que, em relação ao total de fonemas, a ocorrência desses casos de imprevisibilidade da língua portuguesa é baixa.

No mesmo texto, verificou-se que havia 19 ocorrências de /X/, entre 13.157 fonemas analisados, e 10 delas eram nas palavras “peróxido” e “profilaxia”, palavras técnicas da odontologia.

Nesse mesmo relatório de conversão preliminar, checaram-se erros de tradução e de prosódia. Não havia erro de prosódia, o que significa que esse tipo de erro ocorre pouquíssimo no sistema. Os erros de tradução, também poucos, encaixaram-se nas impropriedades antecipadas. Cabe lembrar que esses desvios se devem a falha de programação, e não a arbítrio do computador.

A análise de vários relatórios preliminares de conversão do *corpus* odontológico e de outros textos mostra que, considerando-se que nem todos os textos contêm todos

os itens que o Nhenhém não traduz a contento, o nível de acerto na conversão pode ultrapassar 98% em relação ao sistema alfabético previsível. Os casos de grafia imprevisível não são erros de sistema, e sim irregularidade lingüística, portanto, estão fora da hipótese da pesquisa, em que se considera apenas o que é previsível no sistema escrito padrão do português do Brasil. Ao se considerar o sistema alfabético em seu todo, o nível mínimo de acerto de transcrição do Nhenhém é de 95%.

A partir disso, a funcionalidade do sistema comprova a hipótese de pesquisa, ao permitir alto nível de acerto na descodificação automática grafema-fonema: em torno de 98%. Com mais estudos, esse porcentual tende a aumentar. Ainda, o êxito obtido evidencia que o elevado grau de transparência do sistema alfabético do português do Brasil se deve, em boa medida, a ele fundamentar-se em intuições fonológicas.

4.3 ALÉM DO NHENHÉM

O alto nível de transparência do sistema ortográfico do português do Brasil deve ser usado como ferramenta auxiliar na alfabetização. Por exemplo, o aluno aprende que se escreve “m” antes de “p” e “b”, como se isso fosse uma regra para complicar o português escrito. Ensinar assim demonstra despreparo do professor em relação a seu objeto de ensino. Nesse caso, o aluno deve saber que o motivo para haver essa “regra” vem do respeito à fala. Afinal, se /p/ e /b/ são bilabiais, inevitavelmente, ao preparar a boca para articular esses fonemas, a melhor letra para codificar a nasalização é “m”, porque, como “p” e “b”, “m” é bilabial. Na conversão, as letras “m” e “n” desaparecem quando terminam sílaba, pois, se nasalizam a vogal que as antecede, no lugar delas fica a vogal nasalizada, como em “compra” → /kˈõpra/ e “manto” → /mˈãtu/, afinal, pelo fenômeno da co-articulação, a vogal nasalizada, ao fechar-se, já prepara os articuladores, de acordo com a consoante seguinte.

Se o aluno perceber essa lógica, certamente procurará outras na língua, tentará perceber razões, questionará. Assim, desperta-se a consciência lingüística, que é primordial para superar dificuldades ortográficas e até mesmo o analfabetismo funcional.

É preciso fazer exercícios com os alunos em sala, em que pronunciem as palavras e sintam como os fonemas se convertem na fala, para que tomem consciência de que muitas regras ortográficas obedecem à fala. Então, elas foram feitas para facilitar a leitura e a escrita, e não para dificultá-las, como muitos professores e, conseqüentemente, alunos podem pensar, quando aprendem algo que não lhes é explicado de forma racional.

Para avaliar a competência lingüística do aluno, o professor pode comparar um texto padrão com o mesmo texto, mas escrito pelo aluno. Os relatórios de conversão elucidarão as diferenças entre fonemas e a estatística revelará em traços e números os desvios e acertos que tais textos apresentam. Pode ficar mais fácil lidar com alguns problemas de alfabetização. Por exemplo, ao pedir, oralmente, para um usuário digitar a palavra “ócio” ele digitou “ósseo” no programa. Isso confirmou involuntariamente a coincidência de sons entre vogais agrupadas, que interfere na concepção do usuário sobre o que é dito, pois a transcrição de ambas as palavras é /ˈɔsiu/.

É difícil explicar que uma letra pode assumir mais de um valor, dependendo das letras que a cercam, mais difícil ainda é explicar que uma letra, circundada pelas mesmas letras, pode ter mais de um valor. Isso reafirma a ineficácia de alfabetizar pelo nome das letras, e não pelo valor que elas assumem nos contextos em que figuram (SCLIAR-CABRAL, 2003). Talvez por isso uma pessoa semi-alfabetizada disse à pesquisadora que conhece todas as letras, mas não sabe uni-las. Ora, para uni-las é preciso estar atento às transformações de valores que sofrem e provocam ao juntar-se a outras. Brincando com o Nhenhém, aluno e professor podem ver algumas dessas transformações e constatar como são lógicas e necessárias ou não, ou seja, poderão formar juízo. Poderão conferir as pronúncias, discordar delas. Para isso também é necessário reflexão.

Despertar a consciência do aluno não significa que ele encontrará respostas lógicas para tudo, mas que procurará respostas, entenderá que há motivos para as regras ser como são. Um passo a dar nesse sentido é conscientizar o aluno de Letras, de Pedagogia, os futuros professores, de que ao aluno deve ser dada a opção de conhecer e manipular eficientemente a língua padrão, sem que isso gere discriminação dos

muitos falares brasileiros. Essa é a língua que ele usará para escrever de forma a ser entendido por todos no país. Essa é a língua que ele usará para entender o que está escrito em jornais, revistas, livros. Essa é a língua que ele usará para entender os programas televisivos, os programas sobre agricultura, sobre o campo. Essa é a língua que ele usará para escrever a esses programas, pedindo ajuda e informações sobre situações de sua área.

A função da escola é habilitá-lo para isso, sem agredir variações sociolingüísticas, sem agredir a língua padrão.[43] A escola deve tentar garantir-lhe inserção social, preparação para lidar com situações mais amplas da vida, e não mantê-lo encarcerado em seu mundo, sem iniciativa para melhorar, encarar uma sociedade diferente da sua, a qual lhe pode trazer benefícios e prejudicá-lo.

A língua tem de ser ensinada na escola, e, como anota o lingüista francês Tonnelat (*apud* CÂMARA JR., 1986), o ensino escolar tem de assentar necessariamente numa regulamentação imperativa. Não adianta ignorar isso.

Aliás, a fuga disso tem provocado lacunas danosas ao ensino de língua materna e facilitado o estabelecimento do analfabetismo funcional, sobretudo porque a escola, na falta de o que ensinar e como ensinar em lugar dos padrões, perde-se no conteúdo, e o aluno não aprende nem padrões. Falta bom senso e preocupação com desenvolver melhores formas de ensinar. Falta uma proposta para ocupar o lugar do famigerado e mal falado padrão. Falta reconhecer que não há como evitar as regras, e mesmo a lingüística contemporânea mais radical desdobra-se em busca de padrões para suas novas regras.[44]

Conhecer bem a língua materna é a melhor forma de lidar com estrangeirismos, ou seja, quebras de padrão. O cidadão poderá fazer reconhecimento da palavra, notar incompatibilidades entre ela e o sistema de sua língua materna, enfim, poderá refletir, formar opinião, posicionar-se com segurança.

Espera-se que o Nhenhém contribua nesse sentido, como ferramenta para ajudar a entender a língua escrita, bem como para submetê-la a testes e brincadeiras. Que tal

---

[43] Especificamente sobre esse tema sugere-se ler Vasilévski (2004a).

[44] Sobre questões de lingüística contemporânea associadas ao ensino e uso da norma padrão, conferir Lopes da Silva e Rajagopalan (2004).

fazer reconhecimento das diferenças entre os alfabetos ortográfico e fonológico do português? Que tal achar rimas testando a terminação das palavras? Que tal testar as rimas criadas pelos poetas brasileiros, portugueses e de outros países em que se fala português? Que tal soltar a criatividade?

Relatórios de transcrições do Nhenhém podem auxiliar no trabalho de estudantes e professores de Letras e Pedagogia e de pós-graduação. Os pesquisadores e professores das áreas de fonologia, fonética e alfabetização também podem beneficiar-se da conversão e da estatística do Nhenhém. Eles podem montar massas de dados a partir de textos de alunos, comunidades, áreas, com classificações determinadas, e aplicar estatística especificamente a esse *corpus*. Mesmo a área de literatura pode desfrutar o Nhenhém, na comparação e avaliação, inclusive estatística, de poemas e prosas.

A estatística pode ser analisada em sala de aula, por exemplo, se cada aluno fizer um *corpus* específico sobre uma dada categoria e levar o relatório de estatística para comparação e discussão. O que os números revelam? Haverá diferenças, semelhanças entre os resultados? O que elas indicam? Certamente, as realidades e necessidades respaldarão o uso do aplicativo Nhenhém 1.0.

4.4 O NHENHÉM E O ACORDO ORTOGRÁFICO

Como mencionado, o acordo, apesar de alterar menos a ortografia brasileira do que a portuguesa, trará algumas implicações fonológicas ao programa. Cabe analisar o comportamento do Nhenhém 1.0 frente às novas regras.

O hífen será excluído em alguns casos, como em “anti-religioso” e “contra-senso”. Isso não provocará alterações no programa, ao contrário, facilitará a tradução, porque o usuário digitará apenas uma palavra em vez de duas, e o programa não aceita hífen. Assim, para o programa, a palavra terá apenas um radical: “contra-senso”, “contrassenso” → /kˈõtra sˈẽsu/, /kõtrasˈẽsu/.

A exclusão dos acentos diferenciais dos ditongos abertos “éi” e “ói” tornará imprevisíveis no programa também esses casos de vogais abertas. As traduções de

“idéia” e “bóia” → /idˈɛya/ e /bˈɔya/ passarão a /idˈeya/ e /bˈoya/ na versão 1.0. Palavras como “dêem” e “vôo” → /dˈẽy/ e /vˈou/, ao perderem o acento gráfico, passarão a /dẽy/ e /vow/. A primeira perderá apenas a prosódia no programa, que a entenderá como monossílabo (uma vogal apenas), a segunda virará ditongo, e a tradução ficará totalmente errada.

A exclusão do trema trará a maior perda, pois tornará imprevisível a pronúncia do “u” dos dígrafos “gü” e “qü”. Assim, “agüentar” e “eqüino” → /agwẽtˈaR/ e /ekwˈinu/ passarão a /agẽtˈaR/ e /ekˈinu/.

A sistematização eletrônica de encontros vocálicos é sempre complexa. Testes preliminares de adaptação do programa à nova ortografia revelaram que os dois ditongos abertos (“éi” e “ói”) permanecerão imprevisíveis, a princípio, no entanto, não será difícil acertar a tradução no caso da perda de acento das vogais duplas “êe” e “ôo”, sobretudo por ocorrerem em sílaba final da palavra. Quanto à queda do trema, os mesmos testes mostraram que há como prever por regras alguns casos, outros permanecerão imprevisíveis, a princípio.

A inclusão das três letras poderá acarretar problemas na prosódia, pois não há regras para palavras que terminam com elas. No entanto, as terminações das palavras derivadas de estrangeirismos pertencem ao sistema escrito do português.

5 CONCLUSÃO

uma carta uma brasa através
por dentro do texto
nuvem cheia da minha chuva
cruza o deserto por mim
a montanha caminha
o mar entre os dois
uma sílaba um soluço
um sim um não um ai
sinais dizendo nós
quando não estamos mais

Leminski (2008)

Dado o tempo de pesquisa, o nível de erros e acertos, as possibilidades oferecidas, o sistema de conversão grafema-fonema Nhenhém versão 1.0 obteve êxito como resultado do projeto que originou esta tese, embora provavelmente haja quem o avalie pelos 5% máximos de erro que pode provocar na conversão fonológica, e não pelos 95% mínimos de acerto.

Por o sistema trabalhar maiormente com o fonema, mas a sílaba ser, como defendem alguns autores, a real estrutura elementar da língua, seu acerto de transcrição grafema-grafofonema atinge ótimo percentual, e revela o quanto da língua se pode transcrever considerando-se apenas o fonema. Isso motiva a seguir em frente com a programação, a refazer ou adaptar o sistema com base na sílaba ou intercalá-la nele, para comparar o acerto nas respostas e as novas possibilidades de relatórios numéricos. Um estudo aprofundado da sílaba, com opções de solução para casos controversos, seria necessário para inserir um molde silábico num programa de computador. Nesse caso, estudos específicos em morfologia contribuiriam para incrementar regras no sistema. Quase não há morfologia no Nhenhém.

A prática revela que outras regras podem ser desenvolvidas para dar conta de aspectos peculiares da língua portuguesa do Brasil, sobretudo a partir da tonicidade. Fenômenos de juntura vocabular podem ser inseridos no programa, e isso seria menos complexo do que lidar com vogais abertas não acentuadas graficamente. No entanto,

tais fenômenos implicam reanálise de muitas regras lingüísticas que o sistema comporta, como ocorreu com a prosódia, além disso, cobrir tais fenômenos não era previsto no projeto de doutorado.

A experiência adquirida mostra os graus de dificuldade de leitura e conversão eletrônica da língua. O nível do fonema é o mais fácil de sistematizar, a dificuldade é maior no nível da sílaba, a morfologia vem a seguir e depois a sintaxe, que é o mais intricado. A complexidade de cada nível pode ser amenizada pela sistematização dos níveis anteriores, porque um se vale do outro. Então, estima-se que o Nhenhém seja um passo para futuro trabalho com níveis que transcendem o fonema.

A questão das vogais abertas não marcadas graficamente, ponto em que o programa é mais falho, mas que também, a princípio, não tem previsibilidade sólida, daria outra tese. Estudos preliminares mostraram que há como sistematizar boa parte dessas ocorrências, não somente os casos de certos verbos, a partir da prosódia e da estatística.

Depois de muita pesquisa em lingüística, devem-se arrolar regras encontradas e desenvolver regras de interação entre linguagem de máquina e língua: os algoritmos. A seguir adaptam-se os algoritmos à linguagem de programação, então o maior trabalho começa: desenvolver as regras no programa. O nível de desvio do funcionamento esperado é considerável e, não raro, desanimador. Por isso, o lingüista deve ter boas noções de programação de computadores, ótimo conhecimento da estrutura da língua padrão e persistência. A partir da língua padrão, outras variedades podem ser adaptadas.

Para o Nhenhém, não importa tanto o fato de a língua ter muitas regras, importa sim que elas sejam gerais e que as exceções tenham regras também. Dessa maneira, a grafia do “x”, com seus casos não previsíveis e falhas nos casos previsíveis, configura obstáculo à sistematização eletrônica do sistema alfabético do português do Brasil. Isso reforça o fato de haver dificuldade na reprodução correta dessa letra na ortografia. A questão das palavras iniciadas por “trans”, “sub” e “ob” seguidos de “s” é mais complexa ainda. Em outros casos, o uso da letra “s” é complexo, mas integralmente sistematizável. Também o desaparecimento do “r” e “s” finais de verbos (“tâmu”,

“comê”) na fala encaixa-se nesse nível de dificuldade, mas não se escrevem essas palavras assim na língua padrão.

Cabem algumas considerações, no que tange à descodificação, advindas do Nhenhém, em relação ao acordo ortográfico, já que o acordo leva em conta apenas a codificação. A dificuldade de sistematizar o uso da letra “x”, até nos poucos casos em que ela é previsível, indica que um acordo ortográfico (ou uma nova nomenclatura gramatical) eficiente deveria facilitar a descodificação dessa letra. O acordo que está sendo aprovado prevê a queda dos acentos gráficos dos ditongos abertos, como em “idéia”. Esse acento é uma marca apenas diferencial da língua portuguesa do Brasil, trata-se de uma regra especial, e seu desuso implicará maior dificuldade na descodificação do timbre aberto de “e” e “o”, mas não mudará as regras básicas de acentuação do português brasileiro. Já a retirada do trema deixará indistintos os casos em que o “u” que precede “g” e “q” e antecede “e” e “i” é pronunciado. Será uma perda considerável para a descodificação da língua. Tirar trema, acento de ditongos abertos, inserir hífen onde não há, se isso não afetará a pronúncia, porque os tratados interferem apenas na escrita, torna a língua menos fonética, menos lógica.

O sistema de acentuação do português do Brasil é lógico em grande medida, até mesmo os escassos casos especiais o são, pois qualquer palavra que se digite no Nhenhém é corretamente acentuada, mesmo que não exista na língua, mas obedeça a suas regras. As regras especiais, como a dos ditongos abertos, facilitam a descodificação e são, além de intuitivas, lógicas. Caso contrário, o sistema não daria conta delas.

O acordo ortográfico em discussão, mas já com data para entrar em vigor, prevê a inclusão de “k”, “w”, “y” no alfabeto oficial, que serão usadas em casos especiais. A experiência de criar o Nhenhém mostra que a inclusão dessas letras não é útil ao sistema alfabético do português do Brasil. Elas não têm valores distintos de outras, pois há “c”, “qu” e “q” para “k”, “u” e “v” para “w” e “i” para o “y”. Incluí-las complica a descodificação e reduz o nível de transparência do sistema, torna-o menos econômico, portanto, essa atitude desvia-se de um princípio básico da teoria evolutiva

das línguas de Martinet, segundo o qual toda língua procura garantir a economia de seu sistema e as mudanças realizadas são condicionadas por necessidade de economia.

Configuraria equívoco, por exemplo, mudar a ortografia e inserir “w” e “y” no alfabeto oficial para que fossem usados nos ditongos, a fim de facilitar a descodificação, pois a sistematização dos ditongos no Nhenhém mostra que não há necessidade disso. Ainda, na codificação, a perda seria maior: É com “i” ou com “y”?

Simular um uso para o “k” é mais difícil, pois os casos de correspondência a ele como fonema no português (“c”, “qu”, “q”) são assimilados com certa facilidade nos bancos escolares, portanto, não configuram problemas graves de codificação nem descodificação.

A teoria evolutiva também diz que podem ocorrer lacunas na língua que os falantes tentam preencher inconscientemente. Não parece ser esse o caso, pois essas letras são usadas em nomes estrangeiros dados aos brasileiros recém-nascidos (Wallace, Kelly) ou em nomes com grafias criadas por seus pais (Kamilly, Danniely, Yvete, Érik), até ditadas pela numerologia. Muitas vezes, usam-se geminações também (Allana, Anna) e pronúncias inventadas para “y” e “i”, como “Bryan” e “Brian”, em que tanto “y” quanto “i” seria pronunciado como /ay/ ou algo parecido. Essas grafias empobrecem o sistema verbal, do ponto de vista fonológico. Esses e outros reflexos de estrangeirismos “preocupam um número cada vez maior de pesquisadores ao redor do mundo” (RAJAGOPALAN, 2005, p.135).

Voltando ao caso das três letras, não parece haver razão lógica para sua inclusão nas palavras. Desse ponto de vista, o sistema verbal escrito perde consistência. No entanto, pelo acordo, tais letras seriam usadas apenas em casos especiais, como unidades de medidas e palavras derivadas de nomes próprios estrangeiros. Então, continuarão marginais na língua e não acarretarão maiores dificuldades ao sistema escrito. Esse acordo ortográfico não melhora a correspondência entre o sistema escrito e falado do português, não considera a educação escolar, não se pauta em estudos gramaticais e fonológicos – sua motivação é comercial.

Cabe resgatar uma questão do ensino que a criação do Nhenhém evidencia. Na escola o aluno aprende o sistema de acentuação como regras para complicar, mas o

professor deve lidar com isso mostrando ao aluno que esse sistema facilita a leitura e é lógico e econômico – tanto que foi possível inseri-lo completamente no Nhenhém. O aluno precisa entender a parte do sistema de acentuação que se refere ao que não é acentuado graficamente, pois assim se escreve a maioria das palavras em português. Ensinar somente as regras dos acentos gráficos é negligenciar a maior parte da língua.

Por conta de falhas como essa, ao escrever, o aluno desconsidera o trema, por exemplo, nos vocábulos corriqueiros, porque não lhe foi explicado que o trema é um ótimo tira-dúvidas da língua, quando esse aluno mesmo lê um texto. Situação análoga ocorre com acentos gráficos na codificação, e também inversa, pois não é raro ver em textos acentos gráficos e outros diacríticos em usos não oficiais. O trema deixará de fazer parte da língua sem ter sido compreendido por muitos de seus usuários.

Os princípios desenvolvidos para a criação de *corpus* lingüístico fonológico mostraram-se viáveis e foram úteis na validação do programa. Dessa maneira, relaciona-se a lingüística de *corpus* e a fonologia, diretamente por meio da computação. É tema que merece destaque em pesquisas, pois essa técnica ampliará a abrangência da lingüística de *corpus* e a enriquecerá, voltando-a para um campo promissor, em que é pouco pesquisada. Não obstante, a fonologia e a fonética são as áreas que mais se beneficiam com essa distinta forma de averiguar seu objeto. Somente é proveitoso criar um banco de dados fonológico, se houver um programa que o leia e possibilite sua montagem e seu manuseio. Ao que parece, o Nhenhém é pioneiro nesse sentido.

Por fim, cabe reiterar que fica confirmada – com respaldo da lógica de programação de computadores – a transparência do sistema alfabético do português do Brasil. Ratifica-se sua regularidade e uniformidade, pois o nível de cobertura da língua do Nhenhém 1.0 é de, no mínimo, 95%, sem considerar a estrutura da sílaba. Assim, comprova-se a hipótese de que o alto nível de previsibilidade dos valores que os grafemas do sistema alfabético do português do Brasil têm pode ser reproduzido em um sistema computacional baseado em regras, que faz a conversão grafema-fonema de forma automática.

CONSIDERAÇÕES FINAIS

Um sistema eletrônico é feito de versões. A versão do programa Windows que chegou aos usuários em massa foi a 3.1, e não a 1.0. Depois veio a 3.11, que iniciava trabalho com rede, e finalmente a versão 95, que começava a trabalhar com Internet e embutia o sistema operacional, que até então era o Microsoft Disk Operation System (MS-DOS), que não era gráfico nem colorido, e devia ser instalado previamente na máquina.

O aplicativo Nhenhém pode melhorar em versões posteriores. Apesar disso, o fato de se apresentar uma versão suficiente para ser 1.0, cuja distribuição é gratuita e liberada a todos que desejem usá-la – com cópias exclusivas e licenciadas – é incentivo para que a pesquisa em fonologia e em lingüística de *corpus* cresça e que novas versões do Nhenhém e novos sistemas sejam desenvolvidos, a partir, inclusive, da experiência dos usuários com o próprio Nhenhém. O Nhenhém 1.0 está em processo de registro como marca, no Instituto Nacional de Propriedade Industrial (INPI).

Depois de tantos vaivens, desvios de rota, percalços, desassossegos, mas também, e mais importantes, êxitos, fica a certeza de que não se trata de um caminho tão íngreme, dado que um pesquisador – com a assistência de um programador de computadores e orientação acadêmica em lingüística de *corpus* e fonologia – conseguiu chegar até aqui, em quatro anos. Os conhecimentos reunidos na revisão da literatura e na metodologia, juntamente com a experiência de programação, mostram que vale a pena investir em programas lingüísticos eletrônicos.

Cabe enfatizar que é indispensável que eles sejam populares, pois somente o uso poderá aprimorá-los. Quando se trabalha pelo conhecimento, não se podem omitir resultados ou divulgá-los a restrito número de pessoas, sobretudo se envolvem dinheiro público. É questão de responsabilidade com a comunidade acadêmica e com a sociedade. Professores alfabetizadores mostraram-se animados com o Nhenhém, bem como pesquisadores em fonologia. Então, que ele seja aplicado, que venham críticas e sugestões.

REFERÊNCIAS

ACCESS. **Gerenciador eletrônico de banco de dados**. MSOffice XP. Versão 2003. Microsoft Corp.

ALMEIDA, N. M. **Dicionário de questões vernáculas**. 4.ed. São Paulo: Ática, 2001.

ALMEIDA, N. M. **Gramática metódica da língua portuguesa**. 44.ed. São Paulo: Saraiva, 1999.

AURÉLIO. **Novo Dicionário Eletrônico Aurélio**. Versão 5.0. 3.ed. 1. imp. Editora Positivo, 2004.

AURÉLIO. Dicionário Eletrônico Aurélio Século XXI. Versão 3.0, nov. 1999. Lexikon Informática Ltda. Versão integral de: FERREIRA, Aurélio Buarque de. **Novo Dicionário Aurélio** - Século XXI. Rio de Janeiro: Nova Fronteira, 1999.

BARBETTA, P. A. **Estatística aplicada às ciências sociais**. 2.ed. Florianópolis: Ed. da UFSC, 2000.

BECHARA, E. **Moderna gramática portuguesa**. 19.ed. São Paulo: Cia. Editora Nacional, 1973.

BIBER, D.; CONRAD, S.; REPPEN, R. ***Corpus* Linguistics**: investigating language structure and use. Cambridge University Press, Cambridge, 1998.

BISOL, L. O ditongo da perspectiva da fonologia atual. **Revista Delta**, v.5, n.2, p.1885-224, 1989.

BRASIL. Ministério da Educação. Disponível em: <http://mecsrv04.mec.gov.br/acs/asp/noticias/noticiasId.asp?Id=7290>. Acesso em: nov. 2007.

CAGLIARI, L. C. **Análise fonológica**: introdução à teoria e à prática. Campinas: Mercado das Letras, 2002.

CÂMARA JR., J. M. **Problemas de lingüística descritiva**. 16.ed. Petrópolis: Vozes, 1997.

CÂMARA JR. J. M. **Estrutura da língua portuguesa**. 16.ed. Petrópolis: Vozes, 1986.

CÂMARA JR., J. M. **Para o estudo da fonêmica portuguesa**. 2.ed. Rio de Janeiro: Padrão – Livraria Editora, 1977.

CARVALHO, J. A. **Por que se usa m antes de p e b?**. Disponível em: <http://www.gnfcnf.org.br/site/scripts/acervo/artigos/mpb.asp>. Acesso em: 20 jul. 2008.

CARVALHO, M. M. **Ortografias**. 1996. Disponível em: <http://www.dha.lnec.pt/npe/portugues/paginas_pessoais/MMC/Ortograf.html>. Acesso em: 10 nov. 2007.

CLEMENTS, G. N. The geometry of phonological features. In: **Theory of phonological redundancy**. Cambridge University Press, Cambridge, 1985.

CRUZ, A. **Historiografia da lingüística brasileira**. Projeto Primeira Pessoa do Singular: História e estórias da lingüística brasileira. Departamento de Lingüística. São Paulo: CEDOCH-USP. Disponível em: <http://www.fflch.usp.br/dl/cedoch/1a.pessoa/elia.grafia.html>. Acesso em: out. 2007.

CRUZ E SOUSA, J. **Violões que choram**. Disponível em: <http://kplus.cosmo.com.br/materia.asp?co=37&rv=Literatura>. Acesso em: fev. 08.

DELPHI. **Linguagem de programação**. Borland Inc., v. 7, 2002.
GERBER, R. M. **Informática**: o programa Wordsmith Tools. In: GERBER, R. M. e VASILÉVSKI, V. (Orgs.). **Um percurso para pesquisas com base em *corpus***. Florianópolis: Editora da Ufsc, 2007.
GOMES, N. S. **Observações sobre os clíticos**. <http://www.filologia.org.br/revista/artigo/7(20)08.htm>. Acesso em: jan. 2008.
GRÉGGIO, S.; MERKLE, C.; VASILÉVSKI, V. **Transcrição de 'Ofícios do Conselho Supremo Militar ao Governo da Capitania/Presidente da Província – 1821 a 1827'**. Florianópolis, 1999. Não publicado. O original pertence ao acervo do Arquivo Público de Santa Catarina.
INSTITUTO NACIONAL DE PROPRIEDADE INDUSTRIAL – INPI. **Página eletrônica oficial**. Disponível em: <http://www.inpi.gov.br>. Acesso em: jan. 08.
INTERNATIONAL PHONETIC ALPHABET – IPA. Disponível em: <http://www.sil.org/sil/>. Acesso em: out. 2007.
JAKOBSON, R. **Studies on child language and aphasia**. The Hague: Mouton, 1971.
JORNAL O GLOBO, 15 mar. 2008.
JORNAL O GLOBO, 15 nov. 2007.
LEECH, G. *Corpora* and theories of linguistics performance. In: Jan Svartvik (Org.). **Directions in *corpus* linguistics**. Mouton de Gruyter, Berlim, 1992.
LEMINSKI, P. **50 poemas**. Disponível em <http://www.gropius.hpg.ig.com.br/leminski.htm>. Acesso em: jan. 2008.
LEMINSKI, P. **La vie en close**. Brasiliense: São Paulo, 1991.
LOPES DA SILVA, F. e RAJAGOPALAN, K. (Orgs.). **A lingüística que nos faz falhar**. São Paulo: Parábola, 2004.
MALMBERG, B. A fonética: teoria e aplicações. **Caderno de estudos lingüísticos**, Campinas (25):7-24, jul./dez. 1993. Conferências proferidas na UnB, em 7 e 8/06/1983.
MALHA. **História da língua portuguesa**. Disponível em: <http://lportuguesa.malha.net/content/category/3/74/44/>. Acesso em: set. 2006.
MANNING, C. D. e SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. The MIT Press, Cambridge, 2000.
McENERY T. e WILSON A. ***Corpus* linguistics**. Edinburg University Press, Edinburg, 1997.
MS-DOS. **Sistema operacional para computadores**. versão 6.22. Microsoft Corp., 1990.
NAMIUTI, C. **O fenômeno da interpolação na história da colocação de clíticos em português**. Projeto. Disponível em: <http://www.ime.usp.br/~tycho/papers/namiuti_2001b_proj.pdf>. Acesso em: jan. 08.
NHENHÉM: Sistema de Conversão Grafema-Fonema (Curitiba/PR). Vera Vasilévski. **Programa eletrônico para pesquisa em lingüística**. Versão 1.0. Em registro no INPI, Florianópolis-Curitiba, Brasil, março de 2008.
NOMENCLATURA GRAMATICAL BRASILEIRA – NGB. 1959. Disponível em <http://portrasdasletras.folhadaregiao.com.br/ngb.html> acesso em set. 2004.

NÚCLEO INTERINSTITUCIONAL DE LINGÜÍSTICA COMPUTACIONAL – NILC (Org.). ***Corpus* lingüístico**. Disponível em: <http://www.nilc.icmc.usp.br/nilc/tools/*corpora*.htm>. Acesso em: 9 nov. 2003.

OLIVEIRA, F. A. D. **Processamento de linguagem natural**: princípios básicos e a implementação de um analisador sintático de sentenças da língua portuguesa. Disponível em: <http://www.inf.ufrgs.br/procpar/disc/cmp135/trabs/992/Parser/parser.html>. Acesso em: 04 dez. 2003.

PARADOX. **Gerenciador eletrônico de banco de dados**. Borland Inc., 2002.

RAJAGOPALAN, K. **A lingüística de *corpus* no tempo e no espaço**: visão reflexiva. In: GERBER, R. M. e VASILÉVSKI, V. (Orgs.). Um percurso para pesquisas com base em *corpus*. Florianópolis: Editora da Ufsc, 2007c.

RAJAGOPALAN, K. A geopolítica da língua inglesa e seus reflexos no Brasil: por uma política prudente e propositiva. In: LACOSTE, Y. (Org.). e RAJAGOPALAN, K. **A geopolítica do inglês**. São Paulo: Parábola, 2005.

ROCHA, M. Relações anafóricas no português falado: uma abordagem baseada em *corpus*. **Revista Delta**, São Paulo, v. 16, n. 2, 2000. p. 229-261.

SAID ALI, M. **Gramática secundária e Gramática histórica da língua portuguesa**. 3.ed. Brasília: Editora da UnB, 1964.

SANTOS, D.; BICK, E.; MARCHI, R. e AFONSO, S. **O projeto Floresta Sintá(c)tica**: na sua vertente para estudar a língua portuguesa. In: GERBER, R. M. e VASILÉVSKI, V. (Orgs.). **Um percurso para pesquisas com base em *corpus***. Florianópolis: Editora da Ufsc, 2007.

SARDINHA, T. B. Lingüística de *corpus*: histórico e problemática. **Revista Delta**, São Paulo, v. 16, n. 2, 2000. p.323-367.

SCLIAR-CABRAL, L. **Princípios do sistema alfabético do português do Brasil**. São Paulo: Contexto, 2003.

SILVA NETO, S. **História da língua portuguesa**. 5a. ed. Rio de Janeiro: Presença, 1988.

SINCLAIR, J. ***Corpus*, concordance, collocation**. Oxford University Press, Oxford, 1991.

SUMMER INSTITUTE OF LINGUISTICS – SIL. Disponível em: <http://www.arts.gla.ac.uk/ipa/ipa.html>. Acesso em: out. 2007.

UNIVERSIDADE FEDERAL DO PARÁ – UFPA. **Código ASCII**. Disponível em: <http://www2.ufpa.br/dicas/progra/arq-asc.htm>. Acesso em: out. 2007.

UNICODE. **What is unicode?**. Disponível em: <http://www.unicode.org/standard/WhatIsUnicode.html>. Acesso em: out. 2007.

VASILÉVSKI, V. Aspectos histórico-teóricos da lingüística de *corpus*: surgimento, abandono, levante e uso. In: GERBER, R. M. e VASILÉVSKI, V. (Orgs.). **Um percurso para pesquisas com base em *corpus***. Florianópolis: Editora da Ufsc, 2007a.

VASILÉVSKI, V. Lingüística de *corpus*, lingüística computacional e estatística: trio metodológico. In: GERBER, R. M. e VASILÉVSKI, V. (Orgs.). **Um percurso para pesquisas com base em *corpus***. Florianópolis: Editora da Ufsc, 2007b.

VASILÉVSKI, V. Uso e registro de uso do verbo *arrasar*: contribuições para o léxico. In: GERBER, R. M. e VASILÉVSKI, V. (Orgs.). **Um percurso para pesquisas com base em *corpus***. Florianópolis: Editora da Ufsc, 2007c.

VASILÉVSKI, V. Carta aberta ao professor Rajan. In: LOPES DA SILVA, F. e RAJAGOPALAN, K. (Orgs.). **A lingüística que nos faz falhar**. São Paulo: Parábola, 2004a.

VASILÉVSKI, V. L. Rememoração da língua no Brasil para refletir. **Diálogos**, Revista Científica da Unoesc Xanxerê, Xanxerê, ano 3, n.5, v.II, 2004b.

WINDOWS. **Sistema operacional para computadores**. XP Professional. Versão 5.1.2600. Microsoft Corp., 2002.

WIKIPÉDIA. **Acordo Ortográfico de 1990**. Disponível em: <http://pt.wikipedia.org/wiki/Acordo_Ortogr%C3%A1fico_de_1990>. Acesso em: out. 2007.

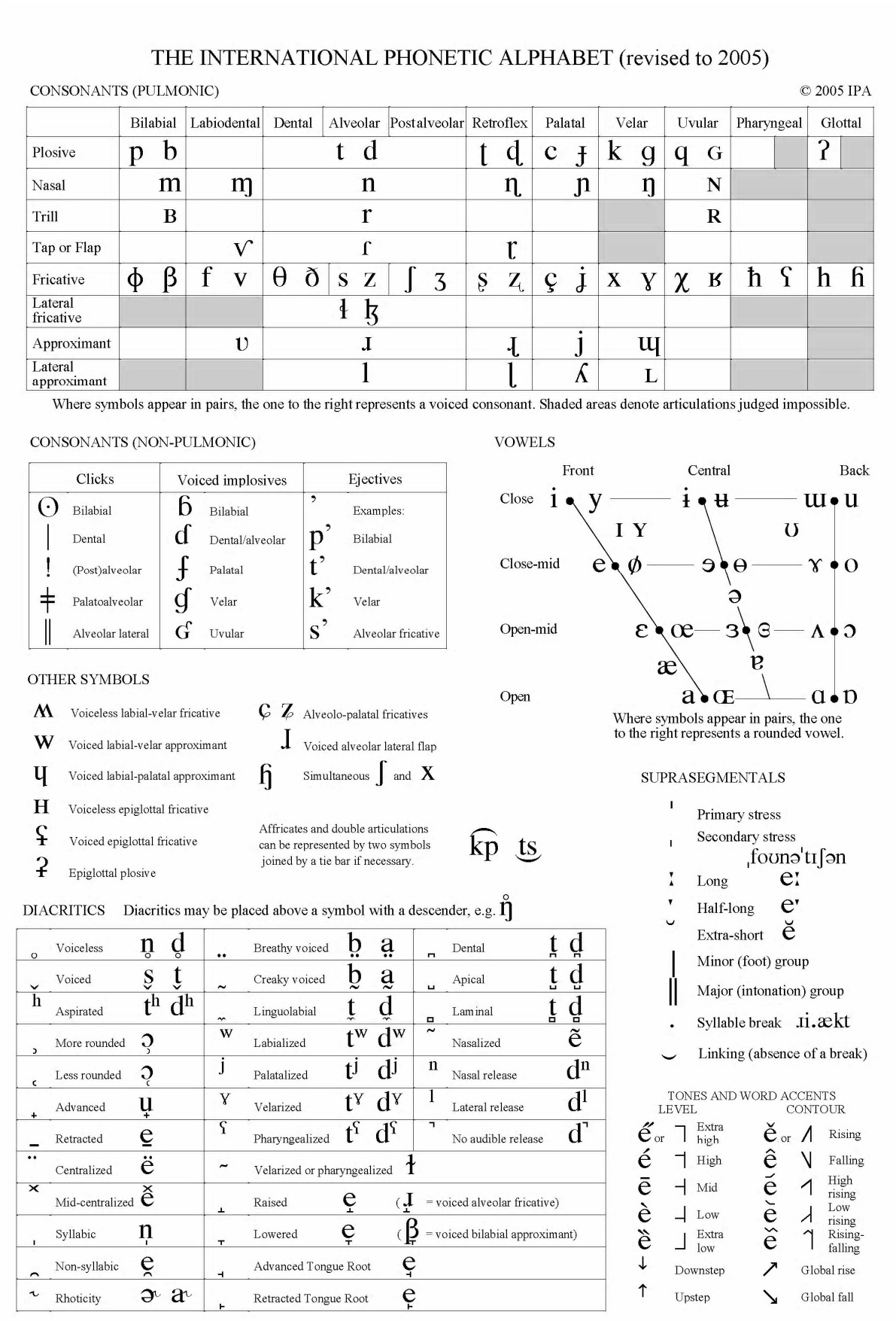
WORD. **Editor de textos eletrônico**. MSOffice XP. Versão 2003 e 2007. Microsoft Corp., 2002, 2006.

WORDSMITH. **Sistema computacional para tratamento de dados lingüísticos**. Versão 3.0. Mike Scott, 1999.

YODA. **Unicode and.NET**. Disponível em: <http://www.yoda.arachsys.com/csharp/unicode.html>. Acesso em: out. 2007.

ANEXOS

ANEXO 1 – ALFABETO FONÉTICO INTERNACIONAL (IPA)



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	’	Examples:
|	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
!	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

VOWELS

Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ	Voiceless labial-velar fricative	ɕ ʑ	Alveolo-palatal fricatives
w	Voiced labial-velar approximant	ɺ	Voiced alveolar lateral flap
ɥ	Voiced labial-palatal approximant	ʧ	Simultaneous ʃ and x
ʜ	Voiceless epiglottal fricative		
ʢ	Voiced epiglottal fricative		
ʡ	Epiglottal plosive		

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.

k͡p t͡s

SUPRASEGMENTALS

ˈ	Primary stress	ˌfoʊnəˈtɪʃən
ˌ	Secondary stress	
ː	Long	eː
ˑ	Half-long	eˑ
˘	Extra-short	ĕ
|	Minor (foot) group	
‖	Major (intonation) group	
.	Syllable break	ɹi.ækt
‿	Linking (absence of a break)	

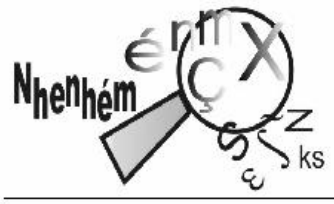
DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. n̥

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
◌ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	◌ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	◌ʲ	Palatalized	tʲ dʲ	◌ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	◌ˠ	Velarized	tˠ dˠ	◌ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	◌ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	◌̴	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
◌̩	Syllabic	n̩	◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
◌˞	Rhoticity	ɚ a˞	◌̙	Retracted Tongue Root	e̙			

TONES AND WORD ACCENTS

LEVEL			CONTOUR	
é᷄ or ˥	Extra high		ě or ˩˥	Rising
é ˦	High		ê ˥˩	Falling
ē ˧	Mid		e᷄ ˧˥	High rising
è ˨	Low		e᷅ ˩˧	Low rising
è᷄ ˩	Extra low		ê᷄ ˥˧˥	Rising-falling
↓	Downstep		↗	Global rise
↑	Upstep		↘	Global fall

ANEXO 2 – RELATÓRIO DE CONVERSÃO GRAFEMA-FONEMA

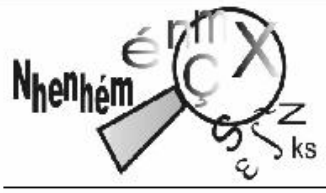


Sistema de conversão grafema-fonema - v.1.0

Relatório preliminar de conversão fonológica

/durˈãti a kaɺˈeyra akadˈemika du mˈɛdiku nuS akoStumˈãmuS a diriʒˈiR nˈosa produsˈãw literˈaria apˈenaS para nˈosuS pˈariS ezˈiSti ũ sẽy nˈumeru di ɺevˈiStaS i lˈivruS tˈɛkinikuS ki diSkˈutẽy uS diferˈẽtiS prosedimˈẽtuS sirˈuRʒikuS utilizˈav eyS nu tratamˈẽtu daS mayS divˈeRsaS kõdisˈõyS klˈinikaS na siruRʒˈia plˈaStika nãw ˈɛ diferˈẽti nu brazˈiw a produsˈãw siẽtˈifika dˈesa eSpesialidˈadi ˈɛ profˈikua ˈoʒi mayS du ki nˈũka aS pesˈoaS si ĩterˈesãw sobremanˈeyra poR tˈudu u ki diS ɺeSpˈeytu ă siruRʒˈia plˈaStika ˈɛ ũ veRdadˈeyru fenˈomenu sosiˈaw maS kwˈãdu si prokˈurãw ˈobraS diriʒˈidaS aw pˈubliku lˈeygu ki sãw uS benefisiˈariuS da sir uRʒˈia plˈaStika i literawmˈẽti a ˈunika ɺazˈãw para sˈua eziStˈẽsia obisˈeRva si grˈ ãdi eSkasˈeS i kwˈãdu ẽkõtrˈadaS mˈoStrãw si frekwẽtemˈẽti demaziˈadu supeRfi siˈayS ˈeSti lˈivru tẽy poRtˈãtu u obiʒetˈivu di foRnesˈeR kõɲesimˈẽtuS sˈobri a sir uRʒˈia plˈaStika aw pˈubliku nãw mˈɛdiku poR mˈeyu di ĩgwˈaʒẽy sˈĩpliS porˈẽy bazeˈada ẽy dˈaduS siẽtˈifikuS uS awtˈoriS diSkˈoɺẽy na fˈoRma di peRgˈũtaS i ɺ eSpˈoStaS sˈobri uS mayS divˈeRsuS prosedimˈẽtuS da siruRʒˈia plˈaStika modˈe Rna a tˈɛkinika operatˈɔria tẽy evoluˈidu kõsideravewmˈẽti naS ˈuwtimaS dˈɛkad aS nãw obiStˈãti sew awkˈãsi nãw ˈɛ ilimitˈadu ˈeSti lˈivru mˈoStra uS diferˈẽtiS aS pˈekituS dˈesa eSpesialidˈadi sirˈuRʒika vˈiStuS di fˈoRma dirˈeta ɺealˈiSta i trãSp arˈẽti sẽy mˈeyaS palˈavraS ẽy sˈuma ˈeSta ˈobra deSnˈuda a siruRʒˈia plˈaStika nu sẽtˈidu di foRnesˈeR subisˈidiuS a tˈoduS akˈeliS ki tˈẽy ĩterˈesi di mayS bẽy kõɲe sˈe la/

ANEXO 3 – RELATÓRIO ESTATÍSTICO FONOLÓGICO



Sistema de conversão grafema-fonema - v.1.0

Relatório estatístico fonológico

TRAÇOS PRINCIPAIS

Silábicos	Assilábicos	Consonantais
5232	420	5252
47,98%	3,85%	48,17%

FONEMAS VOCÁLICOS

Zona de articulação	Timbre	Via de emissão	Movimento dos lábios
Anteriores 2379 42,09%	**Altos** 2503 44,29%	**Orais** 4973 87,99%	**Arredondados** 1713 30,31%
Posteriores 3273 57,91%	**Médios** 1377 24,36%	**Nasais** 679 12,01%	**Distensos** 3939 69,69%
	Baixos 1772 31,35%		

FONEMAS CONSONANTAIS

Modo de articulação			Ponto de articulação	Via de emissão	Fonação
Oclusivos 2725 51,88%			**Anterior** 3339 63,58%	**Orais** 4765 90,73%	**Surdos** 2246 48,95%
Constritivos 2527 48,12%			**Posterior** 877 16,70%	**Nasais** 487 9,27%	**Sonoros** 2342 51,05%
Fricativos 1490 58,96%	**Vibrantes** 784 31,02%	**Laterais** 253 10,01%	**Labial** 1036 19,73%		

PORCENTUAIS E QUANTIAS INDIVIDUAIS DE FONEMAS

/a/ 12% 1339	/ɔ/ 1% 120	/d/ 2% 232	/s/ 4% 453
/ā/ 2% 221	/u/ 7% 732	/t/ 7% 723	/ʒ/ 1% 61
/e/ 5% 588	/ū/ 0% 32	/n/ 2% 197	/ʃ/ 0% 10
/ē/ 2% 243	/y/ 1% 134	/ɲ/ 0% 12	/r/ 4% 484
/ɛ/ 1% 92	/ȳ/ 0% 3	/g/ 0% 38	/Ɍ/ 1% 97
/i/ 11% 1232	/w/ 3% 283	/k/ 4% 435	/l/ 5% 557
/ĩ/ 1% 87	/b/ 1% 102	/v/ 1% 128	/ʎ/ 0% 21
/o/ 4% 421	/p/ 4% 383	/f/ 1% 145	|R| 2% 203
/ō/ 1% 125	/m/ 3% 278	/z/ 2% 232	|S| 4% 461

FONEMAS ANALISADOS: **10904**