Phonologic Patterns of Brazilian Portuguese: a grapheme to phoneme converter based study

Vera Vasilévski
Federal University of Santa Catarina, Emerging Linguistic Productivity Lab (LAPLE), Florianopolis, Brazil
sereiad@hotmail.com

Abstract

This paper presents Brazilian Portuguese phoneme patterns of distribution, according to an automatic grammar rules-based grapheme to phoneme converter. The software Nhenhém (Vasilévski, 2008) was used for treating data: written texts which were decoded into phonologic symbols, forming a corpus, and subjected to a statistical analysis. Results support the high level of predictability of Brazilian Portuguese phonemes distribution, the consonant-vowel syllabic pattern as the most common, as well as the stress pattern distribution 'CV**.**CV#. The efficiency of a phoneme-grapheme converter based entirely on rules is also proven. These results are displayed and discussed, as well as some aspects of Nhe-nhém building.

1 Introduction

The challenging problem of alphabetic systems discovery, i.e., its relationship with the spoken language (Silva Neto, 1988) is the issue discussed, illustrating it with empirical evidence, presenting statistically the Brazilian Portuguese patterns of phoneme distribution, and how they are reflected in the written system. In addition, questions dealing with prosody and syllable are also addressed, with some comments about the spelling agreement that is to be effected in 2013, the goal of which is to standardize the Portuguese spelling in seven countries where it is spoken.

The patterns presented were obtained from the analysis of an automatic grammar rules-based grapheme to phoneme converter designed for dealing with Brazilian Portuguese, the software Nhenhém (Vasilévski, 2008), which is also a syllable parser. The presentation is preceded by a description of the relation between the Portuguese written system and the phonological one and the main problems they cause in finding optimal solutions for writing the program algorithms. Some of the principles of the Portuguese spelling system together with some of the theories that guided the converter construction support the discussion.

2 Spoken and Written Language

Science and also History (Silva Neto, 1988) state that the oral verbal language develops spontaneously whenever traces of humanization are found, whereas the written language is an invention, the intensive and systematic learning of which is necessary in most cases (Scliar-Cabral, 2003a). Linguistic evolution is not just a fact of phonological and phonetic change, however, changes often start as pronunciation modifications (Silva Neto, 1988). Consequently, distinctions fade and disappear, causing homonyms, which must be avoided, so we introduce new words to maintain the independence of signs (Malmberg, 1993). Languages are in perpetual change, although in apparent repose. The distance between the oral and the written system, which is conservative and subject to the literary traditions, becomes increasingly high.

In alphabetic systems, one or more letters (graphemes) represent the phonemes, resulting in units that distinguish meaning in writing (the second articulation), but this representation is not a one-to-one, by virtue of the distance between the oral and the written systems already mentioned. Another divergent principle also occurs: the etymological. Since many spellings are based upon etymological origin (Scliar-Cabral, 2003a) writing does not reproduce the

Proceedings of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss, pages 51–60,
Avignon, France, April 24 2012. ©2012 Association for Computational Linguistics

oral system faithfully. Both spoken and written language have their own laws and ways.

2.1 Phonetics and Phonology

While Phonetics is concerned with describing speech sounds (phones) from the point of view of their articulation, perception and physical properties, Phonology studies the phonemes of a language, that is, classes of sounds, abstractly represented in the minds of a linguistic community. In this way phonemic transcription is broad (general), covering all possible phonetic variations of each phoneme. The aim of Phonology is deep invariance, while Phonetics searches surface variations.

There are many schools of Phonology, the most important of which is the Prague Circle, which introduced the functionalist approach, meaning, in this case, that only phonetic differences which cause differences of meaning are relevant. Perception of those differences is a psychic one and implies disregarding any similar phonetic difference which does not provoke a difference meaning. Phonology makes abstraction of the physical properties of sounds, which are the field of Phonetics. Quoting Glossematics, Phonetics studies the expression of sounds (substance of sounds in their multiplicity and variation), and Phonology studies the form (relations, classes, abstract nature, which takes place in the substance) (Malmberg, 1993).

Since the alphabetic principles are based on the phoneme representation, any automatic program must depart from the phonological description of the respective language, which is the case of the Brazilian Portuguese phonological transcription here used.

2.2 Brazilian Portuguese spelling system

Although the rules of registering stress may seem complicated, they facilitate reading. We will present and discuss here only some of the most important rules regarding the spelling system.[1]

Portuguese is a syllable-timed language, i.e., the vast majority of Portuguese words has stressed syllable, leaving aside clictics, which are only a few, but are the most frequently used (prepositions and accusative pronouns). However, the stressed syllable is not signaled for the most frequent stressed words (the ones which receive stress on the penultimate syllable) since Occam's razor principle was adopted, registering only the stress of less frequent stressed words. The criteria for graphically signaling Portuguese words are the following: a) in which syllable stress falls; b) is it a vowel or consonant that ends the word; c) signaling the difference between diphthong and hiatus.

Signaling graphically stress is a powerful hallmark for the reader, because it guides him/her to match the written word with its representation in the mental oral lexicon. Only meta-language is helpful whenever the diacritic is absent for recovering on which syllable stress falls.

The stress diacritics of Portuguese are acute ("chapéu" – hat) and circumflex ("você" – you). A morphosyntactic diacritic is used for signaling the overlap of the preposition "a" with the definite article "a"/"as", or with the demonstrative pronoun "a"/"aquela(s)", "a"/"aquele(s)". For instance, "fui **à** casa da Maria" (I went to Mary's home), "vamos **à**quele lugar" (Let's go to that place).

In Portuguese, stress may relate to the last, penultimate, antepenultimate or, more rarely, to the fourth last syllable of the phonological word, for example, "núpcias" (wedding) → /'nu.pʲ.si.aS/ (Câmara Jr., 1986). The phonological word in Portuguese is well defined, and its distinctive mark is stress (Câmara Jr., 1986). The stress position reveals, clearly, the distinctive vowel (Câmara Jr., 1997).

The position of stress does not depend on the phonemic structure of the word. There are no word endings in Portuguese imposing certain stress, but there is a termination which is more frequent, although such frequency is indeterminable phonologically (Câmara Jr., 1997). However, the Portuguese characteristic stress occurs in the penultimate syllable, which gives Portuguese a bass rhythm. Nevertheless, Brazilian Portuguese has more words with stress on the last syllable than European Portuguese, because it incorporated words from the African and Indigenous languages that lived together with the Portuguese colonialists in the past.

Portuguese words main stress is registered graphically according to the pattern frequency in the language. The most frequent word pattern is: …'C(C)V.C(C)V(s)#, where the last vowel must be "a", "e", "o". These words do not receive any written signal, e.g., "mesa" (table) → /'me.za/,

[1] Portuguese spelling accent system is showed in details and discussed in Vasilévski (2008).

"escreves" (you write) → /eS.ˈkrɛ.viS/, "livro" (book) → /ˈliv.ru/. Secondly is the pattern …ˈC(C)V(s)#, where the last written vowel must be "a", "e", "o". If the last vowel is [-high, -low], it receives a circumflex, e.g., "avô" (grandfather) → /a.ˈvo/; if the last vowel is [+low], it receives an acute signal, e.g., "sofá" (sofa) → /so.ˈfa/, "cafés" (coffes) → /ka.ˈfɛS/, "vovó" (grandma) → /vo.ˈvɔ/.

On the other hand, the stress of words ending with "i" and "u" – for instance, "abacaxi" (pineapple) and "caju" (cashew) – falls on the last syllable → /a.ba.ka.ˈʃi/ and /ka.ˈʒu/, unless they have accent mark on another syllable, e.g., "júri" (jury),"bônus" (bonus) → /ˈʒu.ri/, /ˈbo.nuS/.

In Brazil, in most of sociolinguistic varieties, the unstressed final vowels spelled with "e" and "o" neutralize in favor of /i/ and /u/, respectively, when pronounced. This neutralization happens because, if the penultimate or antepenultimate syllable of the word is more intense, the last syllable is reduced: "gente" (people) → /ˈʒẽ.ti/, "carro" (car) → /ˈka.ʁu/.

Also, stress of words ending in decrescent diphthongs fall on the last syllable: "plebeu" (commoner) → /ple.ˈbew/, "ramal" (branch) → /ʁa.ˈmaw/, "união" (union) → /u.ni.ˈãw/, unless they have accent mark on another syllable: "pônei" (pony) → /ˈpo.nej /. In Portuguese, all words stressed in the antepenultimate syllable are signaled in writing: "número" (number), "cálida" (warm – fem.), "zênite" (zenith) → /ˈnu.me.ru/, /ˈka.li.da/, /ˈze.ni.ti/.

Another characteristic that makes the Portuguese system of signaling the stressed syllable in the written system effective comes from the fact that it was guided by phonological intuition. One example is a morphosyntactic diacritic exclusive of certain verbs – "ter" (to have), "vir" (to come), and derivatives – in the third person plural ("têm", "vêm") (Scliar-Cabral, 2003a), thus indicating plural, since third person singular is "tem" and "vem"). The pronunciation, however, does not change: "vem", "vêm" → /vẽj/, /ˈvẽj/.

In summary, the Portuguese written system of signaling stress is based on the principle of economy (Occam's razor), considering that the most frequent pattern /ˈCV.CV(s)/ is the one that does not receive a diacritic. Thus, it facilitates decoding, although it may seem more complicated for coding, especially as it is not properly understood by teachers and, therefore, by students. The system has lost some of the qualities based on phonological intuition, due to diachronic changes in the oral system and the lack of spelling rules based on those changes: the 1991 agreement made the situation worse. We will come back to this point.

2.3 The Portuguese syllable

The syllable is the superior unit in which phonemes (vowels and consonants) combine to work on enunciation (Câmara Jr., 1997). Syllable division is deeply studied by Phonology. Its structure types characterize languages. The basic phonemic structure is the syllable, not the phoneme (Jakobson, 1967 apud Câmara Jr., 1986). The syllable in Portuguese can be understood as a set of positions (slope (onset), core (nucleus), and decline (coda)) to be occupied by specific phonemes. The core of the syllable is the only essential position in Portuguese and should be always occupied by a vowel, which is the predominant sound of the syllable. The slope is occupied by consonants and may not be present in the syllable. Further restrictions are made to what may be in decline, which accepts only certain consonants and the semi-vowels /j/, /w/, but can also be empty. In Portuguese the so called free or open syllables, which are the ones that end with a vowel, predominate. This kind of syllables includes simple syllables (V) and open complex (CV). Locked or closed syllables are those ending in consonants (VC, CV(C)C). They are much less frequent in Portuguese, and there are severe constraints, limiting which are the possible consonants in this position (Câmara Jr., 1986).

The most complex syllables in Portuguese are the ones that end with two or three phonemes: CCVVC ("**claus**.tro.fo.bi.a" → /**klawS**.tro.fo.ˈbi.a/), CCVCC ("**trans**.mu.ta.ção" → /**traNS**.mu.ta.ˈsawN/ ~ /**trãS**.mu.ta.ˈsãw/), and CVCCC ("**gangs**.te.ris.mo" → /**gaN.gʲS**.te.ˈriS.mu/ ~ /**gã.gʲS**.te.ˈriS.mu/). In the last two examples, we can see that there can be two phonological interpretations: the first one considers the existence of nasal consonantal coda and disregards the existence of nasal vowels while the second considers the existence of nasal

vowels and the absence of a nasal consonant phoneme in coda position (what the second position admits is the existence of phonetic variants, conditioned by the subsequent consonant). Nhenhém spelling syllable parsing favors the second position. The sequence CCCV is not valid for Brazilian Portuguese. The pronunciation of a foreign word like *stress* is [is.'trɛ.si], so its written form is "es-tresse".

In general, the Portuguese syllable delimitation is clear, but there are three cases where it is floating. There are three groups of vowels contexts in which an unstressed and high vowel may be considered as a semi-vowel, belonging to a diphthong, or as a vowel, forming a hiatus (Câmara Jr., 1997): a) /i/ or /u/ preceded or followed by another unstressed vowel ("var**i**edade", "sa**u**dade", "cu**i**dado"), b) /i / or /u/ followed by a stressed vowel ("p**i**ano", "v**i**ola"), and c) /i/ or /u/ followed by unstressed vowel at the word ending ("índ**i**a", "assíd**u**o"). Phonetically, one can understand these as diphthongs or hiatuses in free variation with no distinctive opposition. Phonologically, however, there is a syllabic not significant variable boundary. In Brazilian Portuguese, they are better understood as hiatus (/va.ri.e.'da.di/, /pi.'ã.nu/, /vi.'ɔ.la/, /'ĩ.di.a/, /a.'si.du.u/), except in the cases in which the second vowel is "i" ou "u", which are better understood as diphthongs: /saw.'da.di/, /kuj.'da.du /.

The above explanation is part of the theory that sustains Nhenhém rules.

3 Methodology, discussion and results

In this section, we present the methodology apllied to the work corpus and the automatic decoder Nhenhém, due to the close relation between them. For the same reason, also we present the results and discuss them.

3.1 The decoder Nhenhém: presentation

The word that gives the program its name, "nhenhém", comes from the Tupi language – spoken by several Indian tribes who lived and continue living in Brazil – and means the endlessly repetition of a movement made by the lips, a sound, as the voice, therefore, an analogue of the word could be "bla, bla, bla".

Nhenhém (/ɲẽ.'ɲẽj/) is a computational program that decodes Brazilian's official writing system into phonological symbols and marks prosody. This program was used for translating, editing, grouping, and searching the work corpus.

What inspired the software development, in 2008, was the high level of transparency of Brazilian Portuguese alphabetic system, although there are some problems, namely the fact the same grapheme "e" or "o" represents respectively two different vowels, /e/, /ɛ/ and /o/, /ɔ/. So, the hypothesis of the availability of the high level of predictability of that system guided the building of a software based on rules, which automatically converted graphemes into phonemes.

Methodologically, the applicative development associates Computational Linguistics, Corpus Linguistics, Statistics, Phonology, and Phonetics. 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 symbols Nhenhém uses for the conversions are displayed in Tab.1.

Graph	Phon	Example
á	/'ã/	**á**gua (water)
à	/ă/	**à**quela (to which)
â	/'ã/	l**âm**pada (light bulb)
ã	/ã/	maç**ã** (apple)
é	/'ɛ/	p**é** (foot)
é	/'ẽ/	cont**é**m (it contains)
ê	/'e/	l**ê**vedo (barm)
ê	/'ẽ/	t**êm**pora, **ên**fase (temple, emphasis)
e	/ɛ/	**e**ra (era)
e	/i/	elefant**e** (elephant)
í	/'i/	l**í**vido (livid)
í	/' ĩ/	l**í**mpido, **í**ndio (clear, Indian)
i	/j/	pe**i**to (breast)
i	/˜j/	mu**i**to (much)
	/ʲ/	ad(**i**)vento (advent)
ó	/'ɔ/	p**ó** (powder)
õ	/õ/	an**õ**es (dwarfs)
ô	/'o/	p**ô**s (it put – past)

ô	/ˈõ/	c**ô**mputo, c**ôn**scio (calculation, conscious)
o	/ɔ/	s**o**mente (only)
o	/o/	c**o**mente (you comment)
o	/w/	mã**o** (hand)
o	/u/	pat**o** (duck)
u	/w/	pa**u**, taq**u**ara (wood, bamboo)
ú	/ˈu/	**ú**til (useful)
ú	/ˈũ/	c**úm**plice, an**ún**cio (accomplice, ad)
ü	/w/	cinq**ü**enta (fifty)
c	/s/	**c**ebola (onion)
c	/k/	a**c**udir (to help)
ch	/ʃ/	a**ch**ar (to find)
g	/ʒ/	**g**ente, a**g**ir (people, to act)
gu	/g/	**gu**erra, **gu**itarra (war, guitar)
h		**h**oje, a**h** (today, oh)
j	/ʒ/	**j**anela (window)
l	/w/	anzo**l** (hook)
l	/l/	**l**ençol (sheet)
lh	/ʎ/	ma**lh**a (mesh)
lh	/l/	fi**lh**inho (sonny)
m	/m/	**m**iar (to meow)
n	/n/	a**n**o (year)
nh	/ɲ/	ni**nh**o (nest)
qu	/k/	**qu**ente, ca**qu**i (hot, khaki)
q	/k/	a**q**uático (aquatic)
r	/r/	ce**r**a, p**r**ata (wax, silver)
r	/R/	amo**r** (love)
r	/ʀ/	mel**r**o, en**r**edo (blackbird, plot)
r	/ʀ/	**r**osto (face)
rr	/ʀ/	ama**rr**ar (to tie)
s	/s/	**s**apo (frog)
s	/S/	mo**s**ca, le**s**ma (fly, snail)
ss	/s/	a**ss**ar (to bake)
sc	/s/	fa**sc**inante (fascinating)
sç	/s/	cre**sç**a (it grows up)
s	/z/	a**s**a (wing)
x	/kʲS/	táx**i** (taxi)
x	/S/	e**x**por (to expose)
x	/z/	e**x**ato (exact)
xc	/s/	e**xc**eção (exception)
z	/z/	a**z**edo (acid)
z	/S/	lu**z** (light)

Table 1: Nhenhém letters, digraphs and corresponding phonemes

3.2 Nhenhém performance

The computational tool we present here is based on rules, i.e., we did not use machine learning based on a training dictionary. Grammatical rules were converted into algorithms and tested within the corpus. A deep and exhaustive study of the grammatical rules that govern the Portuguese written system preceded the design of the tool, consulting the literature on the subject. Internally, the program has all written Portuguese spelling rules (Câmara Jr., 1997, 1986, 1977; Scliar-Cabral, 2003a; Said Ali, 1964; Bechara, 1973; Bisol, 1989; Cagliari, 2002) converted into algorithms, and also the entire Portuguese prosodic system, as it was created by Gonçalves Vianna in 1911, briefly adjusted in 1945 and in 1973 (Bechara, 1973, Scliar-Cabral, 2003a). If the word stress is signaled graphically, the converter reproduces it, if not, Nhenhém applies the spelling rules presented in section 2.2.

Nhenhém bases the translation on a phonologic alphabet, which takes into account the International Phonetic Alphabet (IPA, 2012) fonts, but it gives responses in Arial Unicode MS font (Tab.1). There are no statistics associated to the rules of grammar. We are not worried by the fact that language has many rules: what really matters is that they are general, and that there are rules for the exceptions as well. Unfortunately, some exceptions escape this principle, and became unpredictable, due to the lack of rules. As a result, they are responsible for about 5% or less of Nhenhém translation inaccuracy. We will discuss some of them later.

The software reads relatively huge bunches of data, and bestow phonologic reports with statistical reports. Examining a phonologic corpus rightly assembled, tests done by drawing on the applicative showed that it reaches no less than 98% of accuracy, reproducing the portion of the Brazilian writing system that is predictable by decoding rules. In relation to the written system as a hole, the correctness is not less than

95%. It is known that, to implement the rules in certain groups, it is important to identify the syllabic unit (Almeida & Simões, 2001; Candeias & Perdigão, 2008), however, the first version of Nhenhém (2008) reached at least 95% of accuracy without recognizing the syllabic unit. Such accuracy was measured by testing several texts with the program. This means that, as soon as we approach this issue properly, the results shall become better. Besides this performance, the program also reaches at least 99% of precision at signaling words stress. Such results confirm the hypotheses, and authenticate the high level of predictability of Brazilian alphabetic system, thanks to its phonological basis. It also corroborates that the Brazilian alphabetic system represents the prosody in a logical, accurate, economic and effective manner.

The program does not fulfill some aspects of translating the written texts into phonological transcription, but this happens because there are some exceptions in the Portuguese written system. For instance, in some cases, the letter "x" values are not all predictable by rules. It can be decoded as five different phonemes: /ʃ /, /s/, /z/, /kʲS/, lSl. For example: "graxa", "sintaxe", "exame", "nexo", "texto" → /'gra.ʃa/, /sĩ.'ta.**s**i/, /e.'**z**ã.mi/, /'nɛ.**kʲ**.**s**u/, /'te**S**.tu/. The first two examples represent the unpredictable cases.

There are also some cases of ambiguity, for instance, the letter "s" value after "b", e.g.: "observar" (to observe) → /obʲseR'vaR/, "obséquio" (favor) → /obʲ'zɛkiu/. So, we consider that "s" as representing an archiphoneme: /obʲSer 'vaR/ and /obʲ'Sɛkiu/ (Vasilévski, 2010).

Morphology can also provoke unpredictable situations. For example, the prefix "trans-", which means "accross", causes a pronunciation ambiguity: "transamazônica" (trans+amazônica) is correctly decoded /trã.za.ma.'zo.ni.ka/, but "transiberiana" (trans+siberiana) is decoded */trã.zi.be.ri.'ã.na/ instead of /trã.si.be.ri.'ã.na/, because there is resyllabification. How to instruct a rules-based program that a rule can either be applied or not for the same situation?

This problem can only be solved by associating morphologycal and phonological rules in the program. We approached this issue deeply in a previous work (Vasilévski, 2008). For now, the solution is to edit the translated text so as to correct all these failures.

Furthermore, the vowels [+low] /ɛ/ and /ɔ/ are written "e" and "o", as mentioned, which makes it hard to predict their values, since /o/ and /e/ have the same coding. When they are stressed

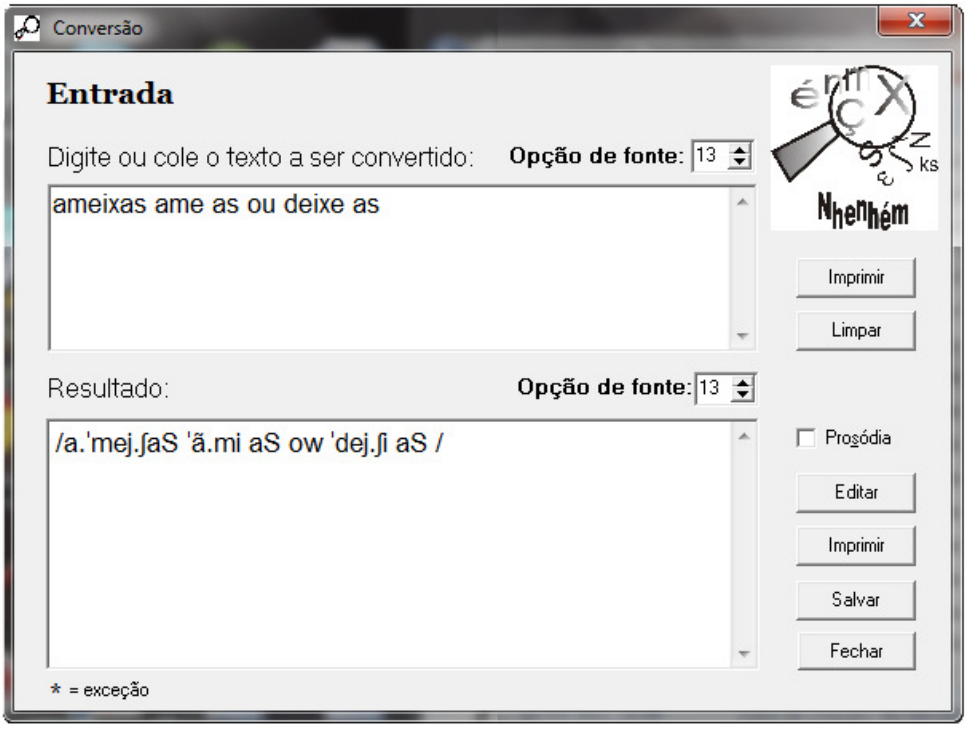

Figure 1: Main screen of the program Nhenhém

and also signaled graphically, the conversion is correct. The reduction of pre-tonic and pos-tonic vowels is also not properly addressed in the Nhenhém algorithm.

Moreover, we decided to consider the so called raising or crescent diphthong as hiatus (Câmara Jr. 1986; Bisol, 1989), therefore words with this ending are decoded as receiving stress on the antepenultimate syllable: "ósseo" → /ˈɔ.si.u/, "história" → /iS.ˈtɔ.ri.a/, "náusea" → /ˈnaw.zi.a/, "ócio" → /ˈɔ.si.u/.

In 2010, Nhenhém was translated into another computer language, and so we could improve its performance. We incremented the main algorithm so that the system became capable of providing the phonological syllabic division, and, consequently, the spelling syllabic division, with at least 99% accuracy. In this way it became easy to signal the stressed syllable, since its 2008 version signaled only the stressed vowel. We used this renewed algorithm to make an automatic syllable parsing for Brazilian Portuguese (Vasilévski, 2010), and we had to solve the problem of syllabication of words that contained hyphen, such as "beija-flor" (hummingbird), "pé-de-moleque" (a peanut candy), "dever-se-ia" (verb to have a duty, conjugated for third person singular, Past Future Indicative, synthetic passive voice, with mesoclisis), and solved them (Vasilévski, 2011).

In addition, we built an interface between Nhenhém and the software Laça-palavras (Vasilévski & Araújo, 2010; Scliar-Cabral & Vasilévski, 2011), which is used for linguistic research. Furthermore, we used the Nhenhém prosodic-phonological algorithm for building a program for speech therapy (Blasi & Vasilévski, 2011), consulting specific literature (Scliar-Cabral, 2003b). This program has been tested and the results were encouraging (Garcez, Blasi, Vasilévski, 2011).

The text is converted while the user types it or pastes it. Pasted texts must have simple formatting, that is, no capital letters. The stressed vowel is signaled by an order from the user. Fig. 1 shows the result for the text "ameixas ame-as ou deixe-as".[2] In the field Resultado (result), the text entry appears converted into phonological symbols. The stressed syllable is signaled by the prosody mark before its first symbol.

[2] Plums love them or leave them – a poem by Paulo Leminski (1991).

The Nhenhém user can automatically convert either one word or a 20 pages text, edit it, save it, research it and print it. As the system conversion is rightly esteemed on at least 95% of accuracy, it allows the user to edit the unsolved 5% (or less) failure rate text, converting, replacing and inserting symbols, adjusting to dialects. The program also allows several texts to be recorded in a database for specific use in statistical reports.

3.3 Phonologic Corpus

In order to test Nhenhém, and also to investigate phonologic patterns of Brazilian written Portuguese, we assembled a corpus with six articles, published in 2007 in a journal of Brazilian dentistry. They are technical and scientific texts, revised, and updated, which were not produced to be used in linguistics research (Sinclair, 1991; Leech, 1992).

The six texts were pre-edited in a text editor, individually, before pasting on Nhe-nhém. Foreign words, words that contained graphemes that do not belong to Portuguese written system and measurement units were eliminated, as well as some acronyms. Some of them could be replaced by its spelling form. The system excludes punctuation, hyphen, quotation marks, and some other symbols by itself, so, they do not need to be treated previously.

In order to reduce chances of conversion errors, care must be taken to ensure the texts' perfect readability by Nhenhém. After this preparation, the corpus texts were pasted on the program, converted, printed, checked, edited, re-checked, and saved for research. The exceptions were searched and edited so as to obtain text correct translations. The texts were loaded for generating statistical reports: the numbers, which will be now exposed, were generated and, as such, are reliable.

3.4 Statistical Report: The Phonologic Patterns

The corpus, after conversion, totalized 69,787 phonemes, being distributed into 33,226 syllabic phonemes (vowels), 3,069 non-syllabic phonemes (semi-vowels), and 33,492 consonant phonemes. Such numbers represent 47.61%, 4.40%, and 47.99% respectively of the total.

To confirm the results, we tested only one of the six texts belonging to the corpus (10,904 phonemes), the numbers of which we present in details (Fig. 2). The main features (traços

principais) distribution is: 47.98% syllabic phonemes, 3.85% non-syllabic phonemes, and 48.17% consonant phonemes. The results are very similar.

Also, the statistical report (Relatório estatístico fonológico) provides phoneme individual distribution, as Tab. 2 displays for the 10,904 phonemes text.

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%		

FONEMAS ANALISADOS: **10904**

Figure 2: Nhenhém statistical report general distribution

In regard to the vowels (fonemas vocálicos), their distribution is: Tongue position: 42.09% front, 57.91% back; Tongue height: 44.29% high, 24.36% mid, 31.36% low; Airstream way (refers to the route taken by the air flow during vocalization): 87.99% oral, 12.01% nasal; Lip rounding: 30.31% rounded, 69.69% unrounded.

The distribution of consonants is: Manner of articulation: 51.88% occlusive, and 48.12% constrictive, distributed as follows: 58.96% fricative, 31.02% vibrating, 10.01% lateral; Place of articulation: 63.58% front, 16.70% back, 19.73% labial; Airstream way: 90.73% oral, 9.27% nasal (oral and nasal); Phonation: 48.95% unvoiced, 51.05% voiced – the archiphonemes |S| and |R| are not included in these numbers, because they neutralize features.

Ph	%	Q		Ph	%	Q
/a/	12,28	1339		|R|	1,86	203
/i/	11,30	1232		/n/	1,81	197
/u/	6,71	732		/f/	1,33	145
/t/	6,63	723		/j/	1,23	134
/e/	5,39	588		/v/	1,17	128
/ l /	5,11	557		/õ/	1,15	125
/r/	4,44	484		/ɔ/	1,10	120
|S|	4,23	461		/b/	0,94	102
/s/	4,15	453		/ʀ/	0,89	97
/k/	3,99	435		/ɛ/	0,84	92
/o/	3,86	421		/ĩ/	0,80	87

/p/	3,51	383		/ʒ/	0,56	61
/w/	2,60	283		/g/	0,35	38
/m/	2,55	278		/ũ/	0,29	32
/ẽ /	2,23	243		/ʎ/	0,19	21
/d/	2,13	232		/ɲ/	0,11	12
/z/	2,13	232		/ʃ/	0,09	10
/ã/	2,03	221		/ɾ̃j/	0,03	3

Table 2: Corpus phoneme individual distribution

A journalistic text composed by 8,454 phonemes was prepared and tested individually by Nhenhém, and the results were similar, since the differences were around 1%. So, the results and also the numbers that show the phonologic patterns of Brazilian Portuguese seem reliable. We tried to find another program or even study that approaches this issue in a similar way, that is, a one that determines the segments from their features and inform such statistics, using corpus, but we did not find any. So, for awhile, we could not make comparisons in order to confirm the reliability of the numbers we have presented.

A lot can be discussed about the results, but we will make general comments here. The back or posterior vowels occur around 15% plus than the front or minus posterior vowels. The posterior ones that appear most are /a/ and /u/, and, among the front, /i/, which occurs only 1% less than /a/. So, the vowel that occurs most in Portuguese is /a/, closely followed by /i/.

The semi-vowel / j̃/ occurs only in the word "muito" (many, much) → /'muj̃.tu/ and derived forms. The /ⁱ/ is computed with /i/, since the first occurs when in a word there is a sequence of two consonants which ordinarily are not a coda, and belong to different syllable. In this case, the epenthetic /ⁱ/ occurs while such sequence is pronounced. So, this inserted phoneme works as core of a phonological syllable: "opção" (option), "cacto" (cactus) → /o.pⁱ.'sãw/, /ka.kⁱ.tu/.

In relation to the consonant phonemes, there is balance in the occurrence of constrictive and occlusive, although occlusive always occur around 3% more than the constrictive ones.

From the results, we find that Brazilian Portuguese phonemic distribution is uniform, once the amount of vowels and consonants tend to be around 50% each. Furthermore, it is possible to deduce that CV (consonant+vowel) is the most common syllable pattern of Brazilian Portuguese. The semi-vowels reveal the amount of diphthongs (the real ones, that is, falling or decrescent diphthongs), since the semi-vowels only occur in this case.

We believe that a deeper analysis of these numbers can be very useful for Portuguese language research.

3.5 The Spelling Agreement of 1991 (2009)

Some changes are to occur in Brazilian Portuguese spelling, due to a spelling agreement, according to which at least seven of the countries where Portuguese is spoken must use the same spelling, from 2013 on.

The most important change for Brazilian Portuguese is the exclusion of the shudder ("trema"), since recognizing diacrisis becomes unpredictable, e.g., the pronunciation of "u" on digraphs "gü" and "qü". Thus, "agüentar" (to stand) and "eqüino" (horse), until 2013 correctly decoded as /agwẽˈtaR/ and /eˈkwinu/, will be spelled "aguentar" and "equino", generating the translations */agẽˈtaR/ and */eˈkinu/. In Brazil, shudder use is still very common. For these reason, Nhenhém will preserve this resource in its algorithm.

This means that the alphabetic system loses transparency, that is, loses one of the rules that make it predictable; therefore, reading (decoding) is impaired. Other changes interfere less in the automatic translation, but none of them disturbs the prosody system.

4 Conclusion and Outlooks

The experience of building, testing and using Nhenhém has shown the degree of linguistic texts electronic reading and conversion difficulty. The phonemic level is the easiest to systematize, the difficulty is greater for the syllable level, the morphology level comes next and then the syntax, which is more intricate. The complexity of each level may be attenuated by the systematization of previous levels, because one takes advantage of the other systematization. So, converters like Nhenhém are a step for future work on levels that transcend the phoneme, like we did to the syllable.

Some decisions taken in the system building are objectionable to some and noteworthy to others, as are some of theories chosen. However, this was not optional. The choices came from the need imposed by the programming and, within

that, objectivity and intelligibility of existing theories, and beliefs and intuition of teachers, students and other language users. The efficiency of Nhenhém confirms the usefulness of the theories adopted.

Now that we have made the automatic syllable parsing, the project follows. We have been working at making the statistical report to look directly to the syllable, and we believe the results will be worthwhile. Some of the next steps are to build a voice synthesizer from Nhenhém, and improve Nhenhém Fonoaud, which is the program for speech therapy. Also, we are working on rules for reducing that 5% (or less) failure rate at the conversion. Since the conversion tool successfully exploits the close correspondence between orthographic representation and pronunciation in Brazilian Portuguese, it can prove to be useful in a range of applications, like in speech therapy.

Acknowledgements

This project is sponsored by CAPES, entity of the Brazilian government for the qualification of human resources, which we thank.

References

Almeida, José João, Simões, Alberto. 2001. Text to speech – A rewriting system approach. Procesamiento del Lenguaje Natural, 27:247-255. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/3366/1854

Bechara, Evanildo. 1973. Moderna gramática portuguesa. 19.ed. Cia. Editora Nacional, São Paulo.

Bisol, Leda. 1989. O ditongo da perspectiva da fonologia atual. Revista Delta, 5(2):185-224.

Blasi, Helena, Vasilévski, Vera. 2011. Programa piloto para transcrição fonética automática na clínica fonoaudiológica. Documentos para el XVI Congresso Internacional de la ALFAL, Universidad de Alcalá, Alcalá de Henares/Madri.

Cagliari, Luiz Carlos. 2002. Análise fonológica: introdução à teoria e à prática. Mercado das Letras, Campinas.

Câmara Jr., Joaquim Mattoso. 1997. Problemas de lingüística descritiva. 16.ed. Vozes, Petrópolis.

Câmara Jr. J. M. 1986. Estrutura da língua portuguesa. 16.ed. Vozes, Petrópolis.

Câmara Jr., J. M. 1977. Para o estudo da fonêmica portuguesa. 2.ed. Padrão, Rio de Janeiro.

Candeias, Sara, Perdigão, Fernando. 2008. Conversor de grafemas para fones baseado em regras para português. In L. Costa, D. Santos, N. Cardoso (Eds.). Perspectivas sobre a Linguateca/Actas do encontro Linguateca: 10 anos, 14, 99-104.

Garcez, Tatiane Moraes, Blasi, Helena Ferro, Vasilévski, Vera. 2011. Aplicação do programa piloto para transcrição fonética automática na clínica fonoaudiológica. Anais do 19º. Congresso Brasileiro e 8º. Congresso Internacional de Fonoaudiologia. São Paulo, Brazil. http://www.sbfa.org.br/portal/suplementorsbfa

International Phonetic Alphabet (IPA). 2012. http://www.langsci.ucl.ac.uk/ipa/ipachart.html

Leminski, Paulo. 1991. La vie en close. Brasiliense, São Paulo.

Leech, Geoffrey. 1992. Corpora and theories of linguistics performance. In J. Svartvik (Org.). Directions in corpus linguistics. Mouton de Gruyter, Berlim.

Malmberg, Bertil. 1993. A fonética: teoria e aplicações. Caderno de Estudos Lingüísticos, 25:7-24.

Said Ali, Manoel. 1964. Gramática secundária e Gramática histórica da língua portuguesa. 3.ed. Editora da UnB, Brasília.

Scliar-Cabral, Leonor. 2003a. Princípios do sistema alfabético do português do Brasil. Contexto, São Paulo.

Scliar-Cabral, Leonor. 2003b. Guia prático de alfabetização. Contexto, São Paulo.

Scliar-Cabral, Leonor, Vasilévski, Vera. 2011. Descrição do português com auxílio de programa computacional de interface. Anais da II Jornada de Descrição do Português (JDP), Cuiabá, Brasil.

Silva Neto, Serafim. 1988. História da língua portuguesa. 5a. ed. Presença, Rio de Janeiro.

Sinclair, John. 1991. Corpus, concordance, collocation. Oxford University Press, Oxford.

Vasilévski, Vera, Araújo, Márcio J. 2010-2012. Laçapalavras: sistema eletrônico para descrição do português brasileiro. LAPLE-UFSC, Florianópolis. https://sites.google.com/ site/sisnhenhem/

Vasilévski, Vera. 2011. O hífen na separação silábica automática. Revista do Simpósio de Estudos Lingüísticos e Literários – SELL, 1(3):657-676.

Vasilévski, Vera. 2010. Divisão silábica automática de texto escrito baseada em princípios fonológicos. Anais do III Encontro de Pós-graduação em Letras da UFS (ENPOLE), São Cristóvão, Sergipe, Brasil.

Vasilévski, Vera. 2008. Construção de um programa computacional para suporte à pesquisa em fonologia do português do Brasil. Tese de doutorado, Universidade Federal de Santa Catarina, Florianópolis, Brasil.