GRAPHEME-TO-PHONE TRANSCRIPTION ALGORITHM FOR TEXT-TO-SPEECH SYSTEMS IN EUROPEAN PORTUGUESE

Daniela Braga
Facultade de Filoloxía da Universidade da Coruña
España
danielabraga@udc.es

Abstract

In this paper, a linguistically rule-based grapheme-to-phone (G2P) transcription algorithm is described for European Portuguese (EP). A G2P, together with the stress determination and the syllable division, is an essential tool in the general architecture of a Text-to-Speech (TTS) system. The G2P is part of the text pre-processing module of the TTS system and its purpose is to convert text into a phonetic transcription that is interpreted by the synthesis engine.

A complete set of phonological and phonetic transcription rules regarding the European Portuguese standard variety is presented. This algorithm was implemented under the C++ framework and tested by using online newspaper articles. The obtained experimental results gave rise to 98,80% of accuracy rate. Future developments in order to increase this value are foreseen. Our purpose with this work is to develop a module/tool that can improve synthetic speech naturalness in European Portuguese. Other applications of this system can be expected like language teaching/learning. These results, together with our perspectives of future improvements, have proved the dramatic importance of linguistic knowledge on the development of TTS.

The present paper is organized as follows: in section 1, it is made the state-of-the-art on this subject and the justification of our approach; in section 2, the annotation conventions are described, the G2P algorithm is presented and some details on the implementation are shown; in section 3, results are discussed and in section 4 some conclusions and future work are presented.

Key Words

Grapheme-to-phone Conversion, Phonological Rules, Speech Processing, Text-to-speech Systems.

Palavras-Chave
Conversão Grafema-fone, Regras Fonológicas, Processamento da Fala, Sistemas de Conversão Texto-fala.

1. **Introduction**

Several frameworks have been proposed to tackle the grapheme-to-phone transcription module of a Text-to-Speech system, among which: information theoretic systems such as decision trees [1], [2] table look-up models [3], dictionary-based approaches [4], [5], linguistically rule-based modules [6], [7], hybrid systems [8], [9], neural networks approaches [10] Finite State Transducers (FST) [11] and statistical approaches [12]. One of the most preferred approaches is dictionary-based, which uses a large dictionary containing the phonetic transcription of a given number of words. This technique has been widely applied to languages whose orthography is roughly phonetically-based, such as English or French. But this approach fails when new words that are not in the dictionary come up.

For European Portuguese language (EP), some synthesizers have been proposed since the early 90's, based on different synthesis techniques, such as the formant-based synthesizer DIXI [13] and the Multivox [14], the articulatory-based synthesizer [15] or the concatenative-based synthesizer [16]. In every case, the G2P is an independent module of the TTS architecture and it is mostly built by using a rule-based approach [14], [17], [18] and a Finite State Transducers approach [19], [20]. However, regarding the rule-based approaches, there seem to be no publication of the complete set of rules for EP available.

There is a large tradition of phonological and phonetic studies in European Portuguese with recognized theoretical and experimental results [21], [22], [23]. However, when regarding the application of this knowledge to TTS systems in EP, a big gap remains to be filled, specifically in what concerns practical rules that allow building a G2P module. As far as we know through publications, this phonological knowledge is only being applied to Brazilian Portuguese (BP) TTS systems [24]. The good stability of EP standard variety helped a lot when deciding the output for each grapheme.

We justify our rule-based approach mainly because: Portuguese is a language

with much phonological regularity; a rule approach is more economic than a phonetic dictionary in terms of space required to store the function; a rule approach is always able to read a new word; a good set of rules based in a linguistic description can cover almost any transcription problem; our rule approach has a very good rate of accuracy and many rules can be adapted to other varieties of Portuguese, such as Brazilian Portuguese and Galician (the language spoken by around 3 million speakers in the north-western province of Spain).

For our algorithm construction, studies on recent EP phonology and phonetics were considered [25], [26], [27] and some of the co-articulation phenomena were already included. Sandhi phenomena are expected to be included in future developments of the system.

2. The Transcription Algorithm

2.1. Annotation Conventions

In Table I, the writing conventions used in the G2P algorithm proposed in section 2.1.2 are presented. Some cases were defined previously in the system, such as the Greek and Latin words and the Masculine and Feminine Pronouns. These phonetic transcriptions that occur in these groups of words also happen in BP standard variety, helping this way to reduce error rates in G2P rule-based modules.

For phonetic annotation, we followed SAMPA transcription symbols [28] with one extension ([l*] to represent the velar lateral consonant which occurs in the EP articulation of the word <sal>, "*salt*").

2.1.1. The Transcription Rules

In the following Tables II to XII, a complete set of transcription rules for every grapheme used in EP is proposed and some examples of occurrence are given. All graphical patterns were considered, including foreign graphemes (such as <k>, <y>, <w>, <ü>), as they may occur in foreign words. Raising and falling diphthongs were also beard in mind, although some of them are incorporated in the rules of graphemes <i> and <u>. Most of the rules presented are stated in literature (e.g. the 26th or 31st rules of grapheme <e>), according to the widely known description of the

phonological rules for the EP standard variety [21], [23]. Many others are stated in normative linguistic literature [25], [29], [30]. Every phonetic output was confirmed by a native Portuguese linguist, this way assuring the correct transcription of all graphic patterns.

#	Symbol	Meaning
1	...	Any grapheme
2	< x >	Grapheme or set of graphemes x
3	[y]	Phone or set of phones y
4	,	Separates options
5	{ x_1, x_2, x_3 }	Set of graphemes
6	< x_1 {x_2, x_3} >	< x_1 x_2 > or < x_1 x_3 >
7	< C / y >	Consonant except <y>
8	< C / {w, z} >	Consonant except < w> and < z>
9	V	Any graphic vowel (e.g. a, e, i, o, u)
10	C	Any graphic consonant (e.g. p, t, k, b, d, g...)
11	Pont	Punctuation mark (e.g. , . ! ? () - ; sp)
12	Ltr	Characters that are letters (e.g. a, b, c, …)
13	SP	Space between words
14	Hf	Hyphen
15	<(case) x >	Certain case that modifies the grapheme <x>
16	<(C) x>	<x> is a consonant
17	<(V) x>	<x> is a vowel
18	<(UV) x>	<x> is unvoiced
19	<(VO) x>	<x> is voiced
20	<(US) x>	<x> is an unstressed vowel
21	<(S) x>	<x> is a stressed vowel
22	<(W_bgn) x>	<x> is in word beginning
23	<(Rad_G) x>	<x> is part of a Greek or a Latin root list (e.g. cardio, homo, macro, meso, micro, otfalm, hiper, etc.)
24	<(Prn_M) x>	Grapheme <x> in Masculine Pronoun and composed words (e.g. este(s), esse(s), aquele(s), neste(s), etc.)
25	<(Prn_F) x>	Grapheme <x> in Feminine Pronoun and composed words (e.g. esta(s), essa(s), aquela(s), ela(s), etc.)

TABLE I: Annotation symbols and conventions used in the G2P transcription for EP.

When conceiving these rules, we tried to design them, as much as possible,

according to graphic patterns, this way reducing their dependence from other modules of text processing, such as the tonic syllable marker and the syllable divisor. Main exceptions to some of the rules are listed below on footnotes.

The transcription of the 26 graphemes that are likely to appear in Portuguese texts is performed by a total amount of 146 rules.

#	Graphical pattern <a>	Phone	Example
1	...<(Rad_G) a>...	[a]	radioterapia
2	... < á, à >...	[a]	rápido
3	... <ão>...	[6~w~]	chão, leilão
4	...<ã>...	[6~]	romã, irmã
5	....<â {m,n}><C/h>...	[6~]	lâmpada
6	... <â> <C>...	[6]	câmara
7	... <am><Pont> ...	[6~w~]	sejam, andam
8	....<a (m,n)><C/h>...	[6~]	campo, canto
9	...<a><l><C/h>...	[a]	calmo, palco
10	...<a><i,u,o><C,Pont>...	[a]	paisagem, ao
11	...<(S) a> <r><Pont>…	[a]	matar, andar
12	...< (S) a><m,n>...	[6]	ramo, banha
13	...<(S) a>...	[a][1]	cacto, gato
14	... < a >...	[6]	amador

TABLE II: Transcription rules for grapheme <a>.

#	Graphical pattern <b>	Phone	Example
1	... <b>...	[b]	albatroz
#	Graphical pattern <c>	Phone	Example
1	...<c><t,ç>...	[ ][2 3]	cacto, acção
2	... < c > < e, i >...	[s]	aceitar
3	...<cc>...	[ks]	occitano
4	... < ç >...	[s]	almoço
5	... < ch >...	[S]	acho
6	... <c>...	[k]	claro

TABLE III: Transcription rules for graphemes <b> and <c> .

[1] The preposition <para> is an exception to this rule.

[2] Exception to this rule: in some words <c> in this context is articulated as [k] (e.g. *bráctea, dicção, facção, facto, ficção, fictício, pictórico, secção, sucção*).

[3] [ ] is a phone that is not pronounced.

#	Graphical pattern <d>	Phone	Example
1	... <d>...	[d]	dote
#	Graphical pattern <e>	Phone	Example
1	... <SP> <e> <SP>...	[i]	Zé e Ana
2	...< (Prn_M) (S) e >...	[e]	ele, este
3	...<(Prn_F)(S) e>...	[E]	ela, esta
4	...< (Rad_G) e >...	[E]	hipertrofia
5	... <õ, ã><e>...	[j~]	põe, mãe
6	...<a><e>...	[j]	Caetano
7	...<ém, em><Pont>...	[6~j~]	alguém
8	…<en><s><Pont>…	[6~j~]	imagens
9	... < é >...	[E]	época
10	...< ê, e> <x> <C>....	[6j]	êxtase
11	....<êm ><Pont>...	[6~j~6~j~]	têm, vêm
12	... <ê {m,n}> < C /h >...	[e~]	ciência
13	... < ê >...	[e]	português
14	...<e><i>...	[6]	aceitar
15	...<e><(ct, cç, cc, gn, pç, pt)>...	[E]	dialecto direcção
16	...<e> <n> <Pont>...	[E]	líquen
17	...<e><m,n><e><Pont>..	[E]	gene, leme
18	...<e {m,n}>< C/h>...	[e~]	lento
19	...<e> <sa, se, ssa, za> <s, Pont>...	[e]	chinesa
20	...<e><la><Pont>...	[E][4]	vela, bela
21	...<e><l><C/h, Pont>...	[E]	sensível
22	...<(W_bgn) e> <s> <C>....	[@]	estrada, esperto
23	...<(W_bgn) e><C, rr, ss> <V>...	[i]	exacto, errado
24	...<(W_bgn) he> <C> ...	[i]	herói herança
25	... < (S) e > <m, n > <V>...	[e]	tema, pena
26	...<(S) e><lh, nh, ch, j>...	[6]	velho, lenha
27	...<(S) e><r><Pont>...	[e][5]	ser, manter
28	...<(S) e><u>....	[e]	meu, deu

[4] The contraction between <por> and the definite article <a> is an exception to this rule. Due to the high frequency of this word and its variations (pelo/a/s) it was implemented as an exception.

[5] This rule is working for verbs, e.g. <colher>[e] (*to harvest*), but not for the homographic nouns, e.g. <colher> [E](*spoon*). Morphological analysis is needed to solve most of these cases.

29	... < (S) e >...	[E]	quero
30	...< (US) e> <o, a>...	[j]	área
31	... < (US) e >...	[@]	índice

TABLE IV: Transcription rules for graphemes <d> and <e>.

#	Graphical pattern <f>	Phone	Example
1	... < f >...	[f]	faca, afiar
#	Graphical pattern <g>	Phone	Example
1	... < g > < e, i >...	[Z]	gelo, giro
2	... < g u > < e, i >...	[g][6]	guindaste
3	... < g >...	[g]	garoto, agora
#	Graphical pattern <h>	Phone	Example
1	... < h >...	[ ]	hoje, hospital

TABLE V: Transcription rules for graphemes <f>, <g> and <h>.

#	Graphical pattern <i>	Phone	Example
1	...<Ltr><i><C/t><i>...	[@][7]	vizinho
2	...< i{m,n}><C/h, Pont>...	[i~]	limbo, sinto, fim
3	...<í{m,n}><**C**/h>...	[i~]	símbolo
4	... <V/{i, u}> <i>...	[j]	coisa, sai
5	...<i e><Pont>...	[i]	superfície
6	... < i, í >...	[i]	líquido, saí
#	Graphical pattern <j>	Phone	Example
1	... < j >...	[Z]	jovem
#	Graphical pattern <k>	Phone	Example
1	... <k>...	[k]	ketchup

TABLE VI: Transcription rules for graphemes <i>, <j> and <k>.

#	Graphical pattern <l>	Phone	Example
1	... < l > <C/h, Pont>...	[l*]	calma, vogal
2	... < l SP> <V>...	[l]	sol a sol
3	... < lh >...	[L]	alho, colher

[6] Exceptions to this transcription can be found in the following words where <gu> is pronounced [gw]: *aguentar, antiguidade, arguente, arguição, arguido, consanguinidade, contíguo, contiguidade, ensanguentar, exiguidade, exíguo, lingueta, linguiça, linguista, pinguim, sagui, saguim, sanguinário, sanguinolento, unguento, unguiforme*

[7] In some words (e.g. *notícia, emitir*) this transcription doesn't occur and <i> is articulated as [i]. A complete list is being buit.

4	...<l>...	[l]	a<u>l</u>i
#	Graphical pattern <m>	Phone	Example
1	... < m >...	[m]	<u>m</u>a<u>m</u>ã
#	Graphical pattern <n>	Phone	Example
1	... <n h>...	[J]	ga<u>nh</u>o
2	... <n>...	[n]	<u>n</u>ata, líque<u>n</u>

TABLE VII: Transcription rules for graphemes <l>, <m> and <n>.

#	Graphical pattern <o>	Phone	Example
1	<Pont><o><Pont>	[u]	<u>o</u> bolo é bom
2	...< (Rad_G) o >...	[O]	hem<u>o</u>globina
3	....<ó>...	[O]	c<u>ó</u>digo, av<u>ó</u>
4	... < õ >...	[o~]	coraç<u>õ</u>es, p<u>õ</u>es
5	...<o{m,n}><C/h, Pont>...	[o~]	c<u>om</u>por, c<u>on</u>de, s<u>om</u>
6	... <ô n ><C/h>...	[o~]	g<u>ôn</u>dola
7	... < ô >...	[o]	s<u>ô</u>frego, p<u>ô</u>r
8	...<ou>...	[o]	<u>ou</u>vir, c<u>ou</u>ve
9	...<o><i>...	[o]	d<u>o</u>is, <u>o</u>ito
10	...<(S) o><r><es, Pont>...	[o][8]	comp<u>or</u>, d<u>or</u>, sab<u>or</u>, sab<u>o</u>res
11	...<o><z><Pont>...	[O][9]	v<u>oz</u>, atr<u>oz</u>
12	...<o><so><Pont>...	[o]	saud<u>o</u>so, caprich<u>o</u>so
13	...<o><sa, sos, sas> <Pont>...	[O]	saud<u>o</u>sa, virtu<u>o</u>sa
14	...<(W_bgn) o><l>...	[O]	<u>o</u>lá, <u>o</u>lhar
15	...<Ltr><o><l><C/h>...	[o]	s<u>o</u>ltar, v<u>o</u>ltar
16	...<o><l><Pont>...	[O]	futeb<u>ol</u>
17	...< (W_bgn) h> <o> <r, s, t>...	[O]	h<u>o</u>rtelã, h<u>o</u>ra, h<u>o</u>stil, h<u>o</u>tel
18	... <(S) o > <m, n>...	[o]	s<u>o</u>ma, s<u>o</u>no
19	...<(S) o><a>...	[o]	perd<u>o</u>a, can<u>o</u>a
20	...<(S) o><o><Pont>...	[o]	v<u>oo</u>, enj<u>oo</u>
21	...<a><o><C/n, Pont>...	[w]	a<u>o</u>, ca<u>os</u>
22	...<(W_bgn)o><C><C>...	[O]	<u>o</u>clusão, <u>o</u>btuso

[8] Exceptions to this rule occur in the following words and their variations in which <o> is pronounced [O]: *maior*, *menor*, *melhor*, *pior*, *suor*, *sénior*, *junior*.

[9] The word <arroz> is an exception: <o> is articulated as [o].

23	... <(US) o >...	[u]	carros, opor
24	...< (S) o >...	[O]	embora, agora

TABLE VIII: Transcription rules for grapheme <o>.

#	Graphical pattern <p>	Phone	Example
1	...<ph>...	[f]	orpheu
2	...<p><t, ç>...	[ ][10]	óptimo
3	... < p >...	[p]	pato
#	Graphical pattern <q>	Phone	Example
1	... < q u >< i, e> ...	[k][11]	quilo, quente
2	... < q >...	[k]	qual, quorum
#	Graphical pattern <r>	Phone	Example
1	... < r r > ...	[R]	carro, arrendar
2	...<(W_bgn) r>...	[R]	rua, rio, rocha
3	... < r >...	[r]	mar, cara

TABLE IX: Transcription rules for graphemes <p>, <q> and <r>.

#	Graphical pattern <s>	Phone	Example
1	... <s h>...	[S]	shiatsu
2	...<(W_bgn) s>...	[s]	saúde, sim
3	... <V> <s> <V>...	[z]	asa
4	...<s><SP><C_VO>...	[Z]	olhos verdes
5	...<s><SP><C_UV>...	[S]	olhos castanhos
6	...<s><SP><V, h>...	[z]	os olhos
7	... < s s > ...	[s]	assar
8	... <s><C_VO>...	[Z]	rasgar
9	... <s><C_UV, Pont/SP>...	[S]	rasca, olhos
10	... <tr(a, â) n> <s> <V>...	[z]	transitar
11	... <ob> <s> <éq>...	[z]	obséquio
12	...<s>...	[s]	cansado

TABLE X: Transcription rules for grapheme <s>.

[10] A complete list of exceptions to this rule in which <p> is pronounced (e.g. *helicóptero, aptidão*) is ongoing work.

[11] Exceptions to this transcription can be found in the following words where <qu> is pronounced [qw]: aquícola, aquista, cinquenta, consequência, delinquência, delinquir, deliquescência, eloquência, eloquente, equestre, equidade, equídio, equidistante, equitativo, exequível, frequência, frequente, obliquidade, querco, quercite, quingentésimo, quinquagenário, quinquagésimo, quiproquó, sequela, tranquilo, tranquilidade, ubiquidade.

#	Graphical pattern <t>	Phone	Example
1	... <t h>...	[t]	Arthur
2	... <t>...	[t]	tacto
#	Graphical pattern <u>	Phone	Example
1	...<m><ui><t>...	[u~j~]	muito
2	... < ü > ...	[w]	lingüística
3	... <g, q> <u><a, o>...	[w]	qual, quorum
4	... <u {m, n} > <C /h, Pont>...	[u~]	abundante, retumbante
5	... <é, e, a, i>< u >...	[w]	céu, seu, caudal
6	... < ú, u >...	[u]	acústica
#	Graphical pattern <v>	Phone	Example
1	... < v >...	[v]	voando
#	Graphical pattern <w>	Phone	Example
1	... <w>...	[w][12]	Watt, workshop

TABLE XI: Transcription rules for graphemes <t>, <u>, <v> and <w>.

#	Graphical pattern <x>	Phone	Example
1	... <(e,ê)> <x> <C_UV>...	[S]	êxtase, texto
2	...<(W_bgn) ine> <x><V>...	[z]	inexorável, inexistente
3	... <(W_bgn) e><x><V>...	[z]	execrar, exame, exagero
4	... <(W_bgn) e> <x> <Hf> <C_VO>...	[Z][13]	ex-marido
5	...<(W_bgn)e> <x> <Hf><C_UV>...	[S]	ex-padre
6	... <(W_bgn) e> <x> <Hf> <V>...	[z]	ex-aluno
7	... <(W_bgn) x>...	[S]	xadrês, xaile
8	...<trou><x>...	[s]	trouxe
9	...<m, pr><o, ó, a, á><x><im>...	[s]	máximo, proximidade
10	...<au><x><íl, il>...	[s]	auxílio
11	....<fl><e,u><x>...	[ks][14]	flexão, reflexo
12	...<ne, fi, se><x>...	[ks][15]	anexo, fixar
13	...<x> <Pont>...	[ks][16]	inox, duplex
14	... <x>...	[S][17]	enxofre

[12] In words of German origin <w> is [v] (e.g. *wagner*, *wagneriano*, *wálchia*).
[13] Exception to this rule is found in the Latin expression "ex-libris" [ks].
[14] Exception to this rule is found in the word <reflexão> articulated with [s].
[15] Exception to this rule is found in the word <fixe> articulated with [S].
[16] Exception to this rule is found in the word <cócci x> articulated with [S].

#	Graphical pattern <y>	Phone	Example
1	... <y> <C> ...	[i]	Yguaçu
2	... <y> ...	[j]	yankee
#	Graphical pattern <z>	Phone	Example
1	... < z SP> <C_UV>...	[S]	faz favor
2	... < z SP> <C_VO>...	[Z]	faz bem
3	... < z SP> < V, h>...	[z]	faz anos
4	... < z > < Pont/ SP>...	[S]	arroz, faz
5	... < z >...	[z]	zumbido

TABLE XII: Transcription rules for graphemes <x>, <y> and <z>.

2.1.2. Implementation of the Transcription Rules

The time needed to generate an acoustic utterance from a text input should be as short as possible. In order to achieve the best performance in the implementation process, we tried to group the rules/patterns according to the type of logical test and common characters. The several tests performed by the system along the decision tree were ordered by the graphemes' frequency of occurrence in the language, which means that it starts by testing the most frequent, leaving the less frequent to the end. Just for reference the following order was considered: <a>, <e>, <o>, <s>, <i>, <r>, <d>, <n>, <t>, <c>, <u>, <m>, <p>, <l>, <v>, <q>, <g>, <h>, <f>, <b>, <j>, <z>, <x>, <k>, <w>, <y>. This order is derived from the analysis of around 1 million characters from the FEUP IPB database [31]. More information and details about graphemes and phones frequencies (and sets of graphemes and phones) in EP can be found in [32]. The frequency of grapheme pairs was also considered. By this way, the system can find a compatible pattern with fewer evaluations. In our system, an average of three tree nodes evaluations is needed for reaching the output pattern.

As much as possible, the rules that were leading to the default output were not implemented. The default output is chosen to be the most frequent among the total outputs for a given grapheme. With this procedure, from the total 146 presented rules only nearly 100 are needed. Temporal performance tests with other G2Ps systems are

[17] Exceptions to this transcription can be found in the following words where <x> is pronounced [ks] in the middle of word: *abnóxio, apoplexia, axial, axila, axiologia, axioma, bissexual, circunflexo, complexão, complexo, conexo, convexão, convexo, crucifixo, filoxera, fixação, fixar, fixo, fluxo, galáxia, hererodoxo, indexação, infixo, inoxidável, intoxicar, íxia, léxico, lexicografia, marxismo, maxilar, maximizar, nexo, nóxio, obnóxio, ortodoxo, oxalato, oxidação, oxidar, oxigénio, oxítono, oxiúro, paradoxo, paralaxe, paroxismo, paroxítono, perplexo, praxis, prefixo, prolixo, proparoxítono, saxofone, sexagésimo, sexagenário*.

foreseen.

Compared with the dictionary based approach, our system requires less memory space and can cope with any new word which makes it a better option for large vocabulary EP TTS. Statistical approaches devise transcription rules from a given set of training utterances, but this usually leads to scattered and complex conditions that should be re-organized. This is a good and universal option when little or no knowledge of the language exists. In our case, the rules were developed using EP linguistic knowledge and 20% of the rules cover 76% of the cases.

3. Results And Discussion

The proposed algorithm was implemented and tested using texts from the daily newspaper *Público* [33] available online. Six articles containing a total number of 10884 characters distributed among 1802 words were used as the text input of our system. The phones originated by our system were manually checked and 98.80% were correctly converted.

Type of error	# occur.	% occur.
Homographs with <e>	21	0.19
Mistake in <a> transcription	1	0.01
Mistake in <e> transcription	26	0.24
Mistake in <o> transcription	31	0.28
Mistake in <u> transcription	2	0.02
Portuguese proper names	2	0.02
Foreign proper names	13	0.12
Foreign words	3	0.03
Toponyms	3	0.03
Words with sub-morphemic alternation	8	0.07
Non implemented exceptions	17	0.16
Acronyms	3	0.03
Total	130	1.19

TABLE XIII: Errors in absolute frequency and percentage relative to total grapheme number.

The errors were marked and classified as shown on Table XIII. From a total of 1.19% of errors (130 wrongly transcribed graphemes), 0.16% are exceptions that were still not implemented for lack of time. Most of the errors come from <e> and <o> grapheme transcription, since these graphemes have a high variability in EP. Many of them occur in homographs (e.g. noun <apelo> [e]/ verb <apelo> [E]) and words with sub-morphemic alternation such as (<l<u>o</u>bos> [o]/ <esp<u>o</u>sos> [O]).

Other mistakes in <e> and <o> grapheme's transcription are due to vocalic alternations which occur along the conjugation of the verbs (e.g. <morr<u>e</u>ram> [e]/ <fiz<u>e</u>ram> [E]). This alternation is well known and can also be predicted by using linguistic information.

Foreign words, proper names and toponyms do represent a problem to be considered, since they sum up a total of 21 mistaken phones.

Comparisons with other G2P modules for Portuguese TTS systems seem to confirm the good performance of our rule-based approach. Besides, our system is more optimized when compared with other similar frameworks, since we only deal with 146 rules, when others need around 600 rules, as stated in [14].

Our good results are very comparable to the G2P for a Brazilian Portuguese TTS system [24], with an accuracy rate of 98.43%. Other rule-based frameworks don't mention accuracy rates [14], [17], [34].

4. Conclusion

In this paper, we described a G2P transcriber linguistically rule-based that is perfectly able to be applied to any EP TTS system. The proposed rules were implemented and tested using newspaper articles from *Público* online. Evaluation of our system resulted in 98.80% of correct transcribed graphemes. Most of the errors were found in homographs, verbal conjugations, foreign words, proper names and toponyms.

From the obtained results, it is our belief that morphological decision in terms of verb/noun will drastically improve the efficiency of our G2P system. This study is already ongoing. Phonological information regarding the verbal flexion of the verbs can also be included in future developments. A careful research on the phonological behavior of the foreign words in EP is also foreseen.

The proposed algorithm can be very useful for other varieties of Portuguese,

and due to this it was adapted and implemented for Brazilian Portuguese as can be seen in another paper also published in the proceedings of this conference [35]. From the overall set of rules presented, 70.5% are fully applicable, 21.9% have the same input pattern but a different output. This means that we have 92.5% of pattern compatibility. We also intend to evaluate the compatibility of our patterns with other varieties of Portuguese.

A comparison of our system with another one using automatically-generated rules obtained by decision trees, for instance, is foreseen.

REFERENCES

[1] Lucassen, J. M.; Mercer, R. L. "Discovering phonemic based forms automatically: an information theoretic approach" in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1984, pp. 42.5.1-42.5.4.

[2] Lucassen, J. M. *Discovering phonemic base forms automatically: an information theoretic approach. Technical Report RC 9833* (#43527), IBM T.J., Watson Research Center, February, 1983.

[3] Bosch, A. and Daelemans, W. "Data-oriented methods for grapheme-to-phoneme conversion". in *Sixth Conference of the European Chapter of the Association for Computational Linguistics (EACL'93)*, 1993, pp. 45-53.

[4] Coker, Cecil H.; Kenneth, W. Church and Liberman, Mark Y. "Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis". in *Conference on Speech Synthesis. European Speech Communication Association*, 1990.

[5] Laporte, E. *Phonetique et transducteurs. Technical report*, Université Paris 7, June, 1993.

[6] Allen, J., Hunnicutt, S. and Klatt, D. *From text to speech : the MITalk system*. Cambridge: Cambridge University Press, 1987.

[7] Kaplan, R. M. and Kay, M. "Regular models of phonological rule systems", in *Computational Linguistics 20(3)*, 1994, pp. 331-378.

[8] Meng, H. M., Seneff, S. and Zue, V. "Phonological parsing for bi-directional letter-to-sound/sound-to-letter generation", in *ARPA Human Language Technology Workshop*, Princeton, USA, 1994.

[9] Huang, C. B., Son-Belt, M. A. and Baggett, D. M. “Generation of pronunciation from orthographies using transformation-based error-driven learning”. in *ICSLP94*,Yokohama, Japan, September 1994.

[10] Sejnowski, T. J. and Rosenberg, C. R. “Parallel networks that learn to pronounce English Text”. In *Complex Systems, 1*, 1987, pp. 145-168.

[11] Roche, E. and Schabes, Y. *Exact Generalization of Finite-State Transductions: Application to Grapheme-to-Phoneme Transcription*. Technical Report TR-95-08, Mitsubishi Electric Research Laboratories, Cambridge, USA, March 1995.

[12] Chotimongkol, A. and Black, A. “Statistically trained orthographic to sound models for Thai”, in *ICSLP2000*, Beijing, China, 2000.

[13] Oliveira, L. C. *Síntese de Fala a Partir de Texto*, Ph.D. dissertation, Instituto Superior Técnico, Universidade Técnica de Lisboa, 1996. Available: http://www.inesc-id.pt/ficheiros/publicacoes/3041.pdf

[14] Teixeira, J. P., Freitas, D., Gouveia, P., Olaszy, G., Németh, G. “Multivox: Conversor Texto-Fala para Português”, in *Proceedings of PROPOR’98*, Porto Alegre, Brazil, November 1998.

[15] Teixeira, A. J. S. *Síntese Articulatória das vogais nasais do Português Europeu*, Ph.D. dissertation, Universidade de Aveiro, 2000.

[16] Barros, M. J. *Estudo comparativo e técnicas de geração de sinal para a síntese da fala*, M.S. thesis, Faculdade de Engenharia da Universidade do Porto, 2001.

[17] Oliveira, L. C., Viana, M. C., Trancoso, I. M. “DIXI - Portuguese Text-to-Speech System”, *Proc. EUROSPEECH'91 - 2nd European Conference on Speech Communication and Technology*, Genoa, Italy, September 1991.

[18] Oliveira, L. C.; Viana, M. C.; Trancoso; I. M. “A Rule-Based Text-to-Speech System for Portuguese”, in *Proc. ICASSP'92 - Int. Conf. on Acoustics, Speech, and Signal Processing*, San Francisco, U.S.A., March 1992.

[19] Caseiro, D. A., Trancoso, I. “Grapheme-to-Phone Using Finite-State Transducers”, in *2002 IEEE Workshop on Speech Synthesis*, September, 2002.

[20] Caseiro, D. A.; Trancoso, I.; Viana, M. Céu; Barros, M. “A Comparative Description of GtoP modules for Portuguese and Mirandese using Finite State Transducers”, in *ICPhS'2003 - 15th International Congress of Phonetic Sciences*, August, 2003.

[21] Mateus, M. H. M.; Andrade, A.; Viana, M. C.; Villalva, A. *Fonética, Fonologia e Morfologia do Português*, Lisboa: Universidade Aberta, 1990.

[22] Martins, M. R. D. *Ouvir Falar. Introdução à Fonética do Português* Caminho: Lisboa, 1998.

[23] Mateus, M. H. M. *Aspectos da Fonologia Portuguesa*. Lisboa: Centro de Estudos Filológicos, 1975 (2ª ed. revista, Centro de Linguística da Universidade de Lisboa, 1980).

[24] Barbosa, F. P.; Resende, F. G.; Gonçalves, C. A., Monserrat, R.; Rosa, M. C. "Grapheme-Phone Transcription Algorithm for a Brazilian Portuguese TTS", in *Proceedings of PROPOR 2003*, June 2003, Faro, Portugal

[25] Cuesta, P.; Luz, M. *Gramática da Língua Portuguesa*. Lisboa: Edições 70, 1971, pp. 281-331.

[26] Rodrigues, M. C. *Lisboa e Braga: Fonologia e Variação*. Lisboa: FCT/MCT, 2003.

[27] Mateus, M.; Andrade, E. *The Phonology of Portuguese*. Oxford: Oxford University Press, 2000.

[28] Cf. http://www.phon.ucl.ac.uk/home/sampa/portug.htm

[29] Estrela, E.; Soares, M. A.; Leitão, M. J. *Saber Escrever, Saber Falar*. Lisboa: Dom Quixote, 2004.

[30] Casteleiro, J. M.(coord.) *Dicionário da Língua Portuguesa Contemporânea da Academia das Ciências de Lisboa*. Lisboa: Editorial Verbo, 2001.

[31] Braga, D.; Freitas, D.; Teixeira, J. P.; Barros, M. J.; Latsh, V. "Phonetic Events from the Labeling the European Portuguese Database for Speech Synthesis, FEUP/IPB-DB", in *Proceedings of Eurospeech 2001 – Scandinavia*, Denmark, September 2001.

[32] Coelho, L. *Etiquetagem Automática de Sinais de Fala – Anotação e Segmentação*, M.S. thesis, Faculdade de Engenharia da Universidade do Porto, 2005.

[33] Available: http://www.publico.clix.pt

[34] Oliveira, C.; Moutinho, L.; Teixeira, A. "Um novo sistema de conversão grafema-fone para PE baseado em transdutores", in *Actas do II Congresso Internacional de Fonética e Fonologia*, Maranhão, Brazil, 2004, accepted for publication.

[35] Silva, D.; Lima, A.; Maia, R.; Braga, D.; Moraes, J. F.; Moraes, J. A.; Resende Jr., F. "A rule-based grapheme-phone converter and stress determination for Brazilian

Portuguese natural language processing", in *VI International Telecommunications Symposium (ITS2006)*, September, 2006, Fortaleza-CE, Brazil.