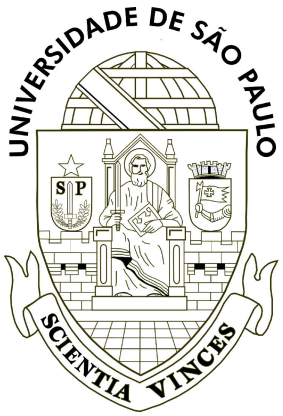
Using phonetic knowledge in tools and resources for Natural Language Processing and Pronunciation Evaluation



Gustavo Augusto de Mendonça Almeida

Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

This dissertation is submitted for the degree of
Master of Sciences

January 2016

To the loving memory of my father,

Tarcízio Otávio Almeida.

1947 – 2003

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Gustavo Augusto de Mendonça Almeida
January 2016

Acknowledgements

And I would like to acknowledge ...

Abstract

Esta dissertação apresenta recursos e ferramentas voltadas para o desenvolvimento de aplicações para reconhecimento de fala, tanto de nativos de não-nativos. São quatro as contribuições aqui discutidas. Primeiro, um conversor grafema-fonema híbrido para o Português Brasileiro, chamado *Aeiouadô*, o qual utiliza regras de transcrição fonética e Classification and Regression Trees (CART) para inferir os fones da fala. Segundo, uma ferramenta de correção automática baseada em aprendizado de máquina, que leva em conta erros de digitação de origem fonética, que é capaz de lidar com erros contextuais e emprega as transcrições geradas pelo *Aeiouadô*. Terceiro, um método para a extração de sentenças foneticamente-ricas, tendo em vista a criação de corpora de fala, baseado em algoritmos gulosos. Quarto, um protótipo de um sistema de reconhecimento e correção de fala não-nativa, voltado para o Inglês falado por aprendizes brasileiros.

Table of contents

List of figures

List of tables

List of acronyms

AmE	American English.
API	Application Programming Interface.
ASCII	American Standard Code for Information Interchange.
ASR	Automatic Speech Recognition.
ATR	Advanced Tongue Root.
BEI	Business English Index.
BP	Brazilian Portuguese.
CAPT	Computer Assisted Pronunciation Training.
CEFR	Common European Framework of Reference.
CG	Computer Graphics.
CMUdict	Carnegie Mellon University Pronouncing Dictionary.
CSR	Continuous Speech Recognition.
CT	Computer Tomography.
DNN	Deep Neural Network.
EF	Education First.
EF-EPI	EF English Proficiency Index.
ESL	English as a Second Language.
F0	Fundamental Frequency.
G2P	Grapheme-to-Phoneme.
GMM	Gaussian Mixture Model.

HDI	Human Development Index.
HMM	Hidden Markov Model.
HSP	Heightened Subglottal Pressure.
HTK	Hidden Markov Model Toolkit.
IPA	International Phonetic Alphabet.
L1	First or Native Language.
L2	Second Language.
MFCC	Mel Frequency Cepstral Coefficients.
MRI	Magnetic Resonance Imaging.
PCM	Pulse Code Modulation.
PER	Phone Error Rate.
PLP	Perceptual Linear Prediction.
POS	Part of Speech.
RASR	RASR.
regex	Regular Expression.
RTF	Real Time Factor.
SNR	Signal-to-Noise Ratio.
UML	Unified Modeling Language.
VBR	Variable Bit Rate.
WER	Word Error Rate.

Chapter 1

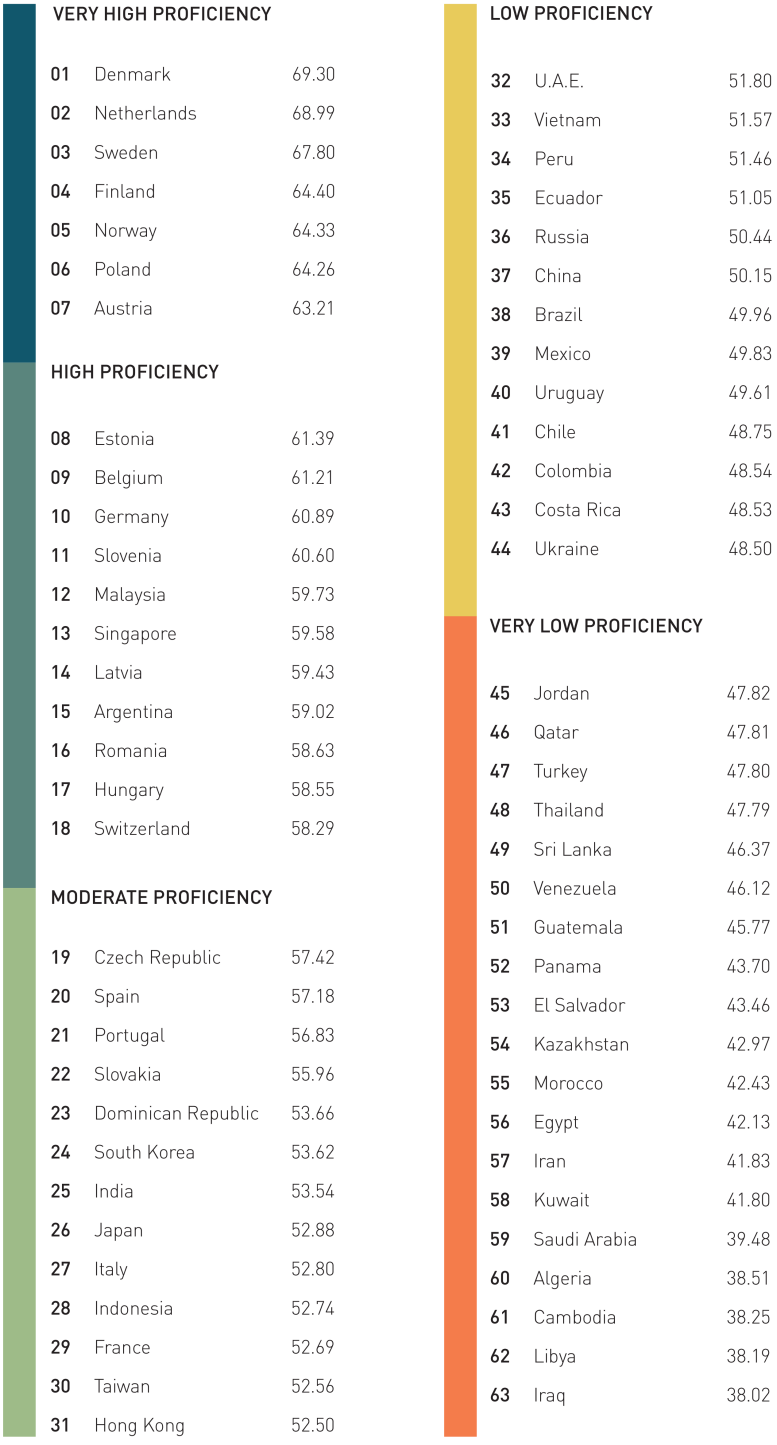
Introduction

Data from the World Economic Outlook, of the IMF (2013), list currently Brazil as the seventh largest economy in the world, with a GDP of US$ 2,396 trillions. According to a survey by The Economist (2013), since 2009, the growth of BRICS (emerging market economies group formed by Brazil, Russia, India, China and South Africa) accounts for 55% of the world economy growth. The current economic scenario is extremely favorable for Brazil to increase its global influence.

However, with regard to the ability to communicate globally, we occupy a much more modest position. In the EF English Proficiency Index (EF-EPI) of 2014, which is published by Education First (EF), Brazil was ranked in the 38^{th} position out of 63 countries, classified among countries with low English proficiency, with 49.96 points [15]. The full ranking is shown in Figure 1.1.

As can be noticed from Figure 1.1, Brazil was immediately behind two other countries from the BRICS, Russia (50.44) and China (50.15); and near several other Latin America countries, such as Peru (51.46), Ecuador (51.05), Uruguay (49.61), Chile (48.75) and Colombia (48.54).

Scandinavian countries lead the very high proficiency rankings, with Denmark (69.30) in the first position, Sweden (67.30) in third the spot, Finland (64.40) in the fourth and Norway (64.33) in the fifth. All these countries have a very high Human Development Index (HDI), according to the United Nations Development Programme. The HDI measures the social and economic development by analyzing three indexes: educational level, average income and longevity. The top five in the EF-EPI rankings are among the top twelve countries in terms of HDI [38]. The EF-EPI data showed that there is a moderate to strong correlation between HDI and English proficiency (R=0.67) [15]. The second position in the EF-EPI ranking is held by the Netherlands, scoring 68.99 points. The Netherlands has the fourth largest HDI



VERY HIGH PROFICIENCY

01	Denmark	69.30
02	Netherlands	68.99
03	Sweden	67.80
04	Finland	64.40
05	Norway	64.33
06	Poland	64.26
07	Austria	63.21

HIGH PROFICIENCY

08	Estonia	61.39
09	Belgium	61.21
10	Germany	60.89
11	Slovenia	60.60
12	Malaysia	59.73
13	Singapore	59.58
14	Latvia	59.43
15	Argentina	59.02
16	Romania	58.63
17	Hungary	58.55
18	Switzerland	58.29

MODERATE PROFICIENCY

19	Czech Republic	57.42
20	Spain	57.18
21	Portugal	56.83
22	Slovakia	55.96
23	Dominican Republic	53.66
24	South Korea	53.62
25	India	53.54
26	Japan	52.88
27	Italy	52.80
28	Indonesia	52.74
29	France	52.69
30	Taiwan	52.56
31	Hong Kong	52.50

LOW PROFICIENCY

32	U.A.E.	51.80
33	Vietnam	51.57
34	Peru	51.46
35	Ecuador	51.05
36	Russia	50.44
37	China	50.15
38	Brazil	49.96
39	Mexico	49.83
40	Uruguay	49.61
41	Chile	48.75
42	Colombia	48.54
43	Costa Rica	48.53
44	Ukraine	48.50

VERY LOW PROFICIENCY

45	Jordan	47.82
46	Qatar	47.81
47	Turkey	47.80
48	Thailand	47.79
49	Sri Lanka	46.37
50	Venezuela	46.12
51	Guatemala	45.77
52	Panama	43.70
53	El Salvador	43.46
54	Kazakhstan	42.97
55	Morocco	42.43
56	Egypt	42.13
57	Iran	41.83
58	Kuwait	41.80
59	Saudi Arabia	39.48
60	Algeria	38.51
61	Cambodia	38.25
62	Libya	38.19
63	Iraq	38.02

Fig. 1.1 English Proficiency Index 2014 Rankings [15].

in the world. This result is similar to that of the previous versions of EF-EPI [14, 13, 12], which also showed a prevalence of Northern European countries in the best positions.

It is interesting to notice that four of the countries at the top five share a common cultural Germanic heritage and speak a language of the Germanic family – the same branch that English is part of. Since Danish, Swedish, Norse, Dutch and English are related through descent from a common ancestor, Proto-Germanic, they inevitably share many linguistic characteristics, such as sound correspondences, a large number of cognates, as well as similar morphology and syntax. This might be an advantage of these countries in comparison to the remaining.

The high proficiency band is occupied mainly by other European countries, both in the Eastern and Western part, such as Estonia (61.39), Belgium (61.21), Germany (60.89), Slovenia (60.60), Latvia (59.43) and Switzerland (58.29). Two Southeast Asian countries also figure within this range, namely Malaysia (59.73) and Singapore (59.58). This result might seem a bit biased since English is one of the official languages in Singapore, along with Malay, Mandarin and Tamil; it is considered the language of business, government, and the medium of instruction in school. We shall highlight Argentina's performance. Despite the great economic depression from 1998 to 2002, Argentina was still able to outperform Brazil, scoring 59.02 points, being the only country from Latin America among the ones with high proficiency.

The moderate proficiency range is filled by the remaining European countries, such as Czech Republic (57.42), Slovakia (55.96), the countries from the Iberian Peninsula – Spain (57.18) and Portugal (56.83)–, together with France (52.69) and Italy (52.80). In addition, the majority of Asian countries which were analyzed in the survey also figure in this list. South Korea achieved the best performance among them, with 53.62 points. India comes next, scoring 53.54 points. The rest of the list is occupied by Japan (52.88), Indonesia (52.74), Taiwan (52.56) and Hong Kong (52.50).

The EF-EPI bands are aligned to the Common European Framework of Reference (CEFR), which is a guideline proposed by the Council of Europe to describe achievements of learners of foreign languages across the European Union. The CEFR reference levels are described in Table 1.1. EF-EPI bands are mapped into CEFR reference levels as follows: the very high proficiency band corresponds to CEFR level B2; very low proficiency to A2; high, moderate and low proficiency bands to B1 with different punctuations.

In case, Brazil's low proficiency rank is analogous to the CEFR B1 level. To put another way, it means that Brazilians are usually able to communicate in English with intermediate skills, being able to understand familiar matters, deal with traveling situations, describe personal experiences and plans, and produce simple texts about subjects of personal interest.

Table 1.1 CEFR reference levels.

Group	Level	Description
Basic User (A)	**Beginner (A1)**	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.
	Elementary (A2)	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
Independent User (B)	**Intermediate (B1)**	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise while traveling in an area where the language is spoken. Can produce simple connected text on topics that are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
	Upper intermediate (B2)	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
Proficient User (C)	**Advanced (C1)**	Can understand a wide range of demanding, longer texts, and recognize implicit meaning. Can express ideas fluently and spontaneously without much obvious searching for express Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organizational patterns, connectors and cohesive devices.
	Proficiency (C2)	Can understand with ease virtually everything heard or read. Can summarize information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in the most complex situations.

As one might observe this is a very restricted communicative competence, which limits English usage basically to the personal domain. To get an idea, the CEFR lists four broad domains: educational, occupational, public, and personal. So Brazilians lacks linguistic abilities in at least three broad domains, this linguistic competence would not allow one to perceive or produce English utterances flexibly, either for social, academic or professional purposes. The performance for each state can be found at the map in Figure 1.2.

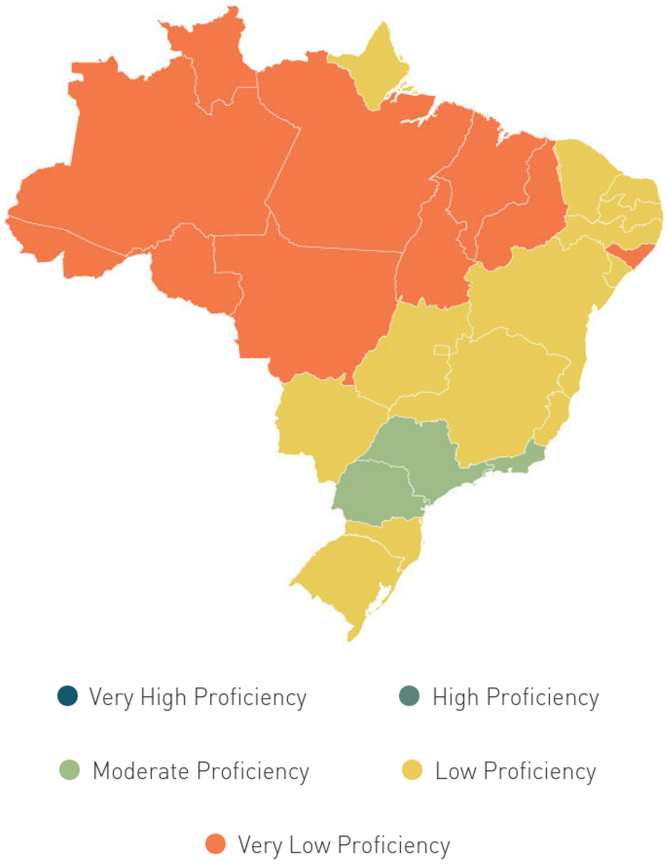


Fig. 1.2 EF EPI scores for each Brazilian state.

As one might see from Figure 1.2, one can clearly see division between Southern and Northern Brazil. Ten states achieved very low proficiency, namely Amazonas, Acre, Pará, Roraima, Piauí, Alagoas, Tocantins, Rondônia, Maranhão and Mato Grosso. The highest scores are found among the richest Brazilian states, mainly in the Southeast and the South regions. The top position is held by São Paulo, which has the largest population, the highest number of industries, and most part of the economic production in the country. Rio de Janeiro, which is the second largest economy of Brazil, occupies the second position. Paraná is the third best ranked state. These three states are the only in the country that achieved

moderate proficiency, they attained a score that is at least 4.8% above the country's average, and all of them are among those with the highest HDI.

With respect of Business English proficiency, our performance is even more concerning. On the 2013 Business English Index (BEI), conducted by GlobalEnglish [18], Brazil reached the 71st position out of 77 countries analyzed. Brazil attained a score of 3.27 points, in a scale from 1 to 10, being placed at the "Beginner" range, the lowest proficiency range considered by the index. Brazil's performance was close to that of El Salvador (3.24), Saudi Arabia (3.14) and Honduras (2.92) which up until recently had experienced civil wars or dictatorship governments. The "Beginner" level is described as: "Can read and communicate using only simple questions and statements, but can't communicate and understand basic business information during phone calls." Again, we can see that this is a very limited linguistic competence, that would not allow one not even to perform the most elementary day-to-day tasks in a company or industry work environment.

Given this scenario, it is clear that we desperately need to improve English language proficiency among Brazilians. This project seeks to be an initial step towards this direction. I developed several tools and resources for non-native speech recognition, focused on Brazilian-accented English or Brazilian Portuguese. In addition, I created a prototype system for non-native speech recognition and correction, called Listener, which is capable of recognizing utterances in Brazilian-accented English and identifying which are the mispronunciations.

Contributions of the Thesis

Within this work, I have investigated and developed a set of tools and resources for non-native speech recognition and correction, focused on Brazilian-accented English. Some of these resources and tools are exclusively for tasks related to processing Brazilian-accented English, but others can also be employed for many other purposes, such as a Grapheme-to-Phoneme (G2P) for Brazilian Portuguese (BP) and balancing of speech corpus. The full list of contributions is provided below:

1. *Aeiouadô G2P*: A grapheme-to-phoneme converter for BP which uses a hybrid approach, based on both handcrafted rules and machine learning method, as decribed in Mendonça and Aluísio [26]. *Aeiouadô dictionary*: A large machine readable dictionary for BP, compiled from a word list extracted from the Portuguese Wikipedia, which was preprocessed in order to filter loanwords, acronyms, scientific names and other spurious data, and then transcribed with Aeiouadô G2P).

2. A phonetic speller for user-generated content in BP, based on machine learning, which takes advantage of Aeiouadô G2P to group phonetically related words, as described in Mendonça et al. [27];

3. A method for the extraction of phonetically rich sentences, i.e. sentences with a high variety of triphones distributed in a uniform fashion, which employs a greedy algorithm for comparing triphone distributions among sentences, as described in Mendonça et al. [28];

4. A crowdsourced platform for speeding up the process of compiling and transcribing speech corpora;

5. A set of rules for generating pronunciation hypothesis for Brazilian-accented English, considering nine types of mispronunciations, respectively: (i) syllable simplification; (ii) consonant change; (iii) deaspiration of voiceless plosives in initial or stressed positions; (iv) terminal devoicing in word-final obstruents; (v) delateralization and rounding of lateral liquids in final position; (vi) vocalization of final nasals; (vii) velar consonantal paragoge; (viii) vowel assimilation; (ix) interconsonantal epenthesis;

6. *Listener*:A prototype system for automatic speech recognition and evaluation of Brazilian-accented English, which makes use of forced alignment, Hidden Markov Model (HMM)/Gaussian Mixture Model (GMM) acoustic models, context free grammars and multipronunciation dictionaries;

All files, resources and scripts developed are available at the project website[1]:

http://nilc.icmc.usp.br/listener

Thesis Structure

This Master's thesis is organized in seven chapters. presents the theoretical fundations, with an introduction to phonetics and phonology, second language acquisition as well as automatic speech recognition. **??** presents th Aeiouadô's grapheme-to-phoneme converter and dictionary. **??** presents a use-case of Aeiouadô, namely a phonetic-speller which employs the transcriptions generated by the grapheme-to-phoneme converter. **??** proposes a method for the extraction of phonetically-rich sentences. **??** describes a prototype system for non-native speech recognition and evaluation of Brazilian-accented English, which makes use of

[1] Due to copyright reasons, the corpora used for training the acoustic models cannot be made available.

the tools and resources developed in this thesis. Finally in chapter 4, we present the overall conclusions, some limitations we found, together with the next steps for future work. A glossary can be found at the back of the thesis in order to help the reader with uncommon terms or specialized jargon.

Chapter 2

Theoretical Foundations

2.1 Phonetics and Phonology

There is an endless debate about what are the boundaries between phonetics and phonology [35]. However, for the purpose of this thesis, we will assume the classical definition, which states that phonetics is the study of the physical properties of the sounds used in languages, whereas phonology is corcerned with how these sounds are organized into patterns and systems [9].

To the first time reader this distinction might seem a bit unclear and confusing. Phonetics main goal is to study the sounds used in speech and provide methods for their description, classification and transcription. On the other hand, phonology is the branch of linguistics which studies sound systems of languages, in other words, how sounds are organized into a system of contrasts which are used distinctively to express meaning [8]. It is interesting to notice that, despite the fact that speech is above all a continuous phenomenon, both phonetics and phonology will conjecture that speech can be examined through discrete units or segments.[1]

Phonetics will analyze the a stream of speech from the viewpoint of a phone, i.e. the smallest perceptible discrete segment in speech [8]. Phones are concrete units, which can be described in terms of their acoustic features or articulatory gestures. Usually, phones are represented with symbols from the International Phonetic Alphabet (IPA), which enrolls all sounds that the human vocal tract could possibly produce. For convenience, the IPA chart is plotted in Figure 2.1.

[1] In this case, we are referring to classical phonetics and phonology. There are contemporary frameworks, such as articulatory phonology or dynamic models, which add time to the equation and consider speech as a continuous phenomenon. But this is beyond the scope of this thesis.

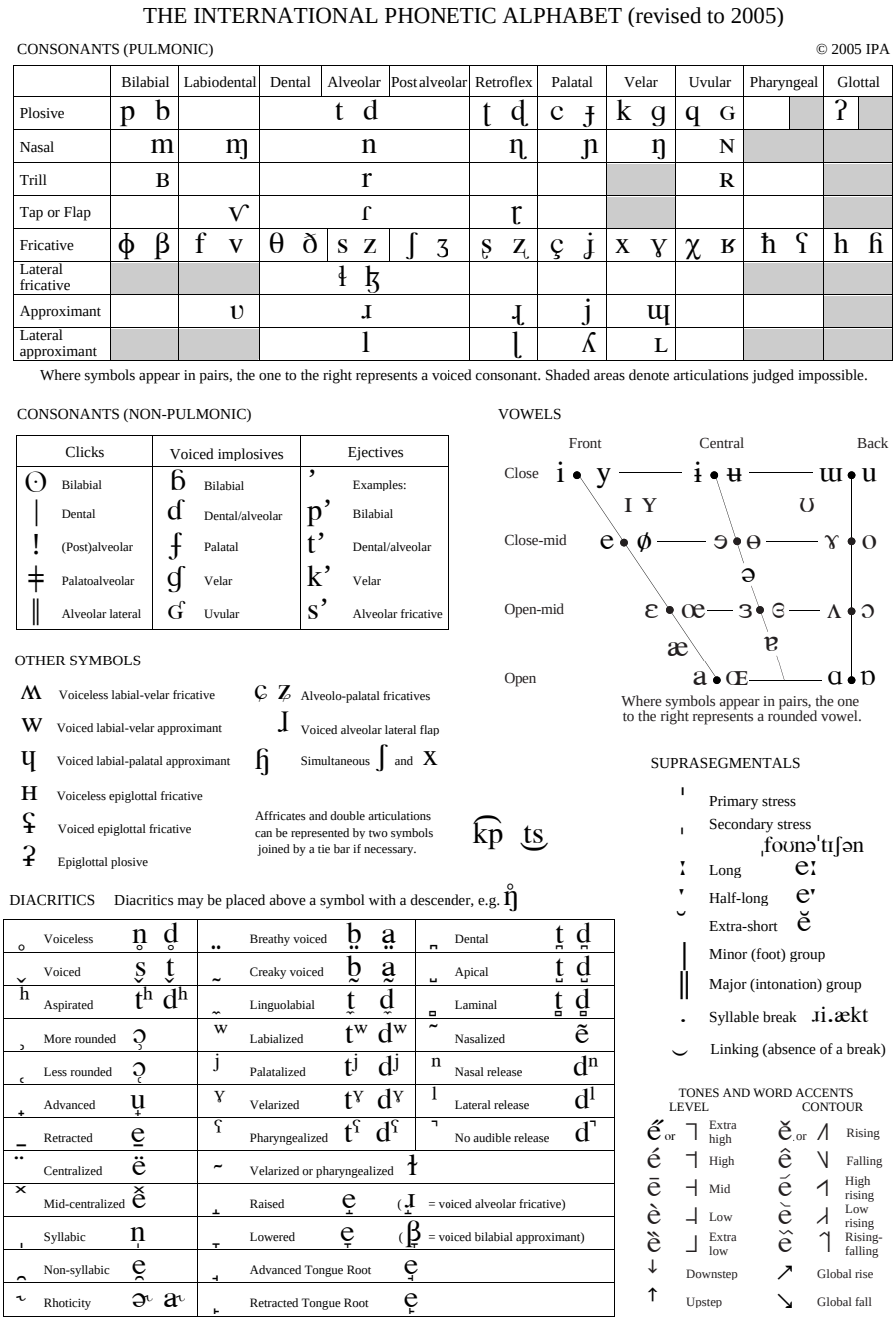


Fig. 2.1 IPA Chart.

The phones in the IPA chart are organized into tables which take into account several properties of the sounds, such as major classes (e.g. “pulmonic consonants” or “vowels”); manner of articulation (e.g. “plosive”, “nasal” or “trill”); place of articulation (e.g. “bilabial”,

"dental" or "alveolar"); status of the glottis (e.g. "voiced", "voiceless" or "aspirated"); type of stress (e.g. "primary" or "secondary"); as well as some other segmental or supra-segmental aspects. For English and BP, the most relevant tables are the ones which contain pulmonic consonants, the table at the top, and vowels, the diagram in the center right position.

Pulmonic consonants are organized as follows: rows designate the manner of articulation, i.e. how the consonant is produced; and columns describe the place of articulation, i.e. where in the phonatory system tract the consonant is articulated. Each cell in the table may contain up to two phones, those which are aligned to the left are devoiced (meaning that the glottis is open when they are produced); and those which are aligned to the right are voiced (which means that the glottis is closed when the phone is uttered).

One refers to each phone by describing its phonetic properties, for instance, the first phone in the table is [p], a voiceless bilabial plosive. It means that the symbol [p] corresponds to a consonant which is produced with a movement of both lips, with the glottis open, in a plosive manner. In other words, [p] describes the sound that is made by first blocking the airflow with both lips closed so that no air can pass, and then by increasing the pressure inside the vocal tract in such way that the air pressure is so high that it bursts the region where it was blocked and passes through, producing sound.

The voiced counterpart of [p] is [b], a voiced bilabial plosive, which means that [b] is produced in the same way of [p], except that for [b] the glottis is closed and not open when the air bursts through the lips. To give a few more examples of how symbols are referred to: [n] is called an alveolar nasal, [ʃ] is a voiceless postalveolar fricative, [ɦ] is a voiced glotal fricative and so on.

Vowels, on the other hand, are described with a different set of features. The vowel diagram (also called vowel trapezium) provides an schematic arrangement of the vowels which summarizes the vowel height of the tongue and/or jaw, as well as how far back the tongue is for articulating each vowel. The vertical position indicates the vowel height, which is related to how close the tongue is to the roof of the mouth or how open is the jaw. Close vowels, which are produced with tongue close to the roof of the mouth, such as the [u] in ***u**va* (grape), are placed at the top of the diagram. In contrast, open vowels, i.e. those which are pronounced with the jaw open or with the tongue distant from the roof of the mouth, such as the [a] in ***a**ve* (bird), are at the bottom of the vowel trapezium. The horizontal position reveals the vowel backness, or the place of the tongue relative to the back of the mouth. Front vowels, such as [i] as in *p**i**pa* (kite), are found in the left part of the vowel diagram; whereas back vowels, like [ɔ] in *ro**ç**a* (small farm), are on the right side.

Vowels and consonants are put together in sequence in order to form words, phrases and sentences. For instance, the word *exceção* (exception) can be transcribed as as the

sequence of phones [e.se'sɐ̃ʊ̃]. As one might notice, the digraph "xc" and the c-cedilla will be mapped into the phone [s], since both graphemes refer to the same sound: a voiceless alveolar fricative. Since in Portuguese we use a script that is quite transparent in terms of letter-to-sound conversion, we tend to assume a one-to-one relation between the number of letters in a word and the number of phones it contains, but this is not always true. For instance, the word *táxi* (taxi) has four letters, but five phones: ['tak.sɪ]; in contrast, *aqui* (here) has four letters but only three phones [a'ki]. Despite their close relation, one must not mistaken letters and phone symbols, the former refers to written language and the latter to the speech stream.

2.1.1 The Phonetic Inventory of Brazilian Portuguese

There is much debate about which set of phones best describes the phonetic inventory of BP. Several analyses have been proposed by different researchers through the years [2, 4, 5, 33, 29], and despite the fact that the analyses usually concur with respect to core questions, there is a lot disagreement in terms of convention and the usage of different phones. For instance, some authors propose that the postonic "a" should be transcribed as [ɐ], whereas others argue that it is more centralized and closer to the schwa [ə]. Similarly, some researchers defend that the glides in Portuguese have a stronger consonantal aspect, thus being transcribed [w] and [j]; at the same time, others argue for a more vocalic nature of these sounds and prefer to represent them as [o̯] and [i̯] respectively.

There is not even a consensus as to which dialect one refers to when one says "BP". As a matter of fact, BP is the native language of nearly 190 million speakers in Brazil [19] and several dialects are currently spoken in different parts of the country. Researchers have different opinions as to what should be considered the standard dialect or the most neutral one.

For the sake of this thesis, we will stick to the analysis put forward by Silva [33], since it is widely known and well-established in the area. Silva [33] proposes 46 phones for describing BP (26 consonants and 20 vowels)[2], all segments are grouped into Table 2.1, Figure 2.2 and Figure 2.3.

As one might notice from Table 2.1, there are six plosive consonants in BP, namely [p, b, t, d, k, g]. As previously said, plosive sounds are produced by first blocking the airflow with both lips closed so that no air can pass, and then by increasing the pressure inside the vocal tract in such way that the air pressure is so high that it bursts through, creating sound. Plosive sounds are also called "stops" or "occlusives". In BP plosive sounds usually occupy

[2]For simplicity, symbols with optional secondary articulation [lˠ, ɫʲ] or with alternative notations [ỹ] were omitted.

Table 2.1 Brazilian Portuguese consonants.

	Bilabial	Labiod.	Alveolar	Postalv.	Palatal	Velar	Glottal
Plosive	p b		t d			k g	
Affricate			tʃ dʒ				
Nasal	m		n		ɲ		
Trill			r				
Tap			ɾ				
Fricative		f v	s z	ʃ ʒ		x ɣ	h ɦ
Approximant			ɹ		j	w	
Lateral Appr.			l		ʎ		

Fig. 2.2 Brazilian Portuguese oral vowels.

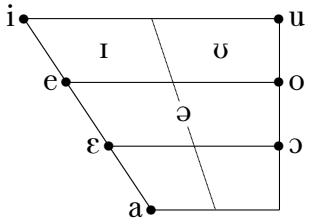
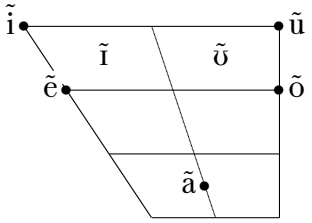


Fig. 2.3 Brazilian Portuguese nasal vowels.



the onset position of a syllabe (i.e. the initial position) as the [p] in ***p****ato* (duck). Some other examples can be found in Table 2.2:

Table 2.2 Examples of plosive consonants in Brazilian Portuguese (I).

Phone	Transcription	Word	Translation	Description
[p]	[p]ato	pato	duck	voiceless bilabial plosive
[b]	[b]ato	bata	(I) hit	voiced bilabial plosive
[t]	mo[t]o	moto	bike	voiceless alveolar plosive
[d]	mo[d]o	modo	way	voiced alveolar plosive
[k]	[k]ato	cato	duck	voiceless velar plosive
[g]	[g]ato	gato	cat	voiced velar plosive

Plosives in BP might also occur in coda position (i.e. the end of a syllable), for instance, as in the [p] *a[p.]to* (able.MASC). However, when plosives occupy coda position in BP

epenthesis will often take place, giving rise to a new syllable structure: a[.pɪ.]to [6]. A few other examples are shown in Table 2.3:

Table 2.3 Examples of plosive consonants in Brazilian Portuguese (I).

Phone	Transcription	Word	Translation	Description
[p]	[ap.]∼[a.pɪ.]to	apto	able.MASC	voiceless bilabial plosive
[b]	[ab.]∼[a.bi.]dicar	abdicar	to abdicate	voiced bilabial plosive
[t]	[at.]∼[a.ti.]∼[a.tʃi.]mosfera	atmosfera	atmosphere	voiceless alveolar plosive
[d]	[ad.]∼[a.di.]∼[a.dʒi.]ministrar	administrar	to manage	voiced alveolar plosive
[k]	[fik.]∼[fi.ki.]ção	ficção	fiction	voiceless velar plosive
[g]	[dɔg.]∼[dɔ.gɪ.]ma	dogma	dogma	voiced velar plosive

BP also has two affricate sounds, both are produced in the postalveolar region: [tʃ] and [dʒ]. Affricate sounds are those that begin by completely stopping the airflow then suddenly releasing it in a constricted way. To put another words, affricates begin with a stop and then are released with a fricative sound, e.g. [tʃ] has two stages, it starts with a [t] stop and then the air is set free with a fricative sound [ʃ].

Affricate phones are often positional variants of [t] and [d], when these are followed by the high vowels [i, ɪ, ĩ], or when occupy coda position. For example, in several dialects of BP, the word *tia* (aunt) is realized as [ˈtʃiə] with an initial devoiced postalveolar affricate [tʃ]. Similarly, *dia* (day) is often pronounced as [ˈdʒiə]. This phenomenon is called palatalization and it results from an overlap among the speech gestures for [t, d] and high vowels [i, ɪ, ĩ]; basically these consonants change their place and manner of articulation in order to anticipate the gestures which are necessary for producing those high vowels.

When [t] and [d] are produced as [tʃ] and [dʒ] due to the presence of a high vowel, they are called positional variants or allophones. Even though in BP [tʃ] and [dʒ] are mainly allophones, there are a few cases when they are not, as the "tch" in ***Tch**u**tch**uca* (pussycat) or the "dj" in ***Dj**avan* (a personal name). It is worth pointing out that in a few dialects of BP, palatalization is a much broader phenomenon which affects other contexts as well [7]. A few other examples of words with affricates are provided in Table 2.4.

Table 2.4 Examples of affricate consonants in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[tʃ]	[tʃi]a	tia	aunt	voiceless postalveolar affricate
[tʃ]	[atʃ.]mosfera	atmosfera	atmosphere	voiceless postalveolar affricate
[tʃ]	[tʃ]u[tʃ]uca	tchutchuca	pussycat	voiceless postalveolar affricate
[dʒ]	[dʒi]a	dia	day	voiced postalveolar affricate
[dʒ]	[adʒ.]ministrar	administrar	to manage	voiced postalveolar affricate
[dʒ]	[dʒ]avan	Djavan	personal name	voiced postalveolar affricate

There are three nasal consonants in BP, viz. [m, n, ɲ]. Nasal consonants are produced with the velum low, in a way that the air is free to pass through the nose. In current language usage, due to vowel nasalization, nasal consonants in BP are basically limited to syllable initial position, for example, as the "m" in *mar* (sea), the "n" in *não* (no) or the "nh" in *rainha* (queen) respectively. It is important to notice that in words such as *ambos* (both) or *anta* (tapir), nasalization most of the time will take place. It means that the gesture for lowering the velum will happen during the articulation of the vowel, in such way that the vowel will be entirely nasalized and the nasal consonant will not perceived as a segment [11], i.e. *ambos* will be produced as [ˈã.bʊs] and *anta* will become [ˈã.tə] with no explicit nasal consonant. A few examples of BP words with nasal consonants can be found in Table 2.5, we also provide some counter-examples of vowel nasalization.

Table 2.5 Examples of nasal consonants and nasalized vowels in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[m]	ca[m]a	cama	bed	bilabial nasal
[n]	ca[n]a	cana	sugar cane	alveolar nasal
[ɲ]	ba[ɲ]a	banha	fat	palatal nasal
(no nasal cons)	[ã]tônio	Antônio	personal name	nasalized [a]
(no nasal cons)	l[ẽ]brar	lembrar	remember	nasalized [e]
(no nasal cons)	[ĩ]teresse	interesse	interest	nasalized [i]
(no nasal cons)	[õ]bro	ombro	shoulder	nasalized [o]
(no nasal cons)	[ũ]tar	untar	grease	nasalized [u]

The sounds [r, ɾ, ɹ, x, ɣ, h, ɦ] are called rhotics because they represent sounds which are somehow related to the letter "r" – "rho" in Greek. Although some of these sounds are quite different in terms of phonetics, phonologically they have shown to behave similarly in many languages [40].

The first one, called alveolar trill [r] is found in some dialects of BP – especiallyin the southern Brazil – and is also known as rolled-r. The alveolar trill is produced by making the tip of the tongue touch the alveolar ridge repeatedly, interrupting the airflow. This sound is part of the rhotics (i.e. the r-like) and for the dialects which have it, it corresponds, for instance, to the "r" in *carta* (letter) or the "rr" in *carro* (car).

The alveolar trill [r] is closely related to the alveolar tap [ɾ], the only difference is that the flap touches the gum ridge once, whereas the trill does it several times. This distinction is found in Spanish, e.g. in *perro* [pɛ.ro] vs. *pero* [pɛ.ɾo]. However, different from the trill, the tap [ɾ] is present in all dialects of BP. It occurs basically in two contexts, between two vowels, e.g. *arara* (parrot), or in complex onsets, such as "br" in *cabrita* (female goat).

The trill is closely related to the alveolar tap [ɾ], the only difference is that the flap touches the gum ridge once, whereas the trill does it several times. This distinction is found in

Spanish, e.g. in *perro* [pɛ.ro] vs. *pero* [pɛ.ɾo]. However, different from the trill, the tap [ɾ] is present in all dialects of BP. It occurs basically in two contexts, between two vowels, e.g. *arara* (parrot), or in complex onsets, such as "br" in *cabrita* (female goat).

The alveolar approximant [ɹ] is the sound which corresponds to the so-called "r-caipira" in BP. It is approximant consonant, which means that the vocal tract is narrowed, but the level of constriction is not sufficient to generate hiss or turbulance. In the dialects in which this rhotic sound occur, its distribution is limited to the end of syllable, as the "r" in *amor* (love) or *porta* (door)

The other rhotics variants [x, ɣ, h, ɦ] can be considered free variants or free allophones amongst themselves, they are also reffered to as strong-r, in contrast to the tap. The first two, [x, χ] are velar fricative sounds consonants, in other words, they are produced in such way that the airflow passes through the vocal tract with constriction and turbulance and their place of articulation is near the soft palate. The phone [x] corresponds to a voiceless sound, which means that the air passes freely through the vocal cords, i.e. they are open. On the other hand, [χ] is a voiced velar fricative, which means that it puts the vocal cords to vibrate when it is produced. The phones [h, ɦ] are articulated in the region of the glottis and they also show constriction in the air passage, that is why the are called glottal fricatives. Analogously to [x, χ], [h, ɦ] also present the voiceless-voiced dichotomy; the vocal cords are open when [h] is produced, but they are closed and vibrate in [ɦ]. Table 2.18 presents some examples of words with rhotic sounds in BP.

Table 2.6 Examples of rhotics in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[r, x, ɣ, h, ɦ]	[r, x, ɣ, h, ɦ]ato	rato	mouse	strong-r
[r, x, ɣ, h, ɦ]	[r, x, ɣ, h, ɦ]oma	Roma	Rome	strong-r
[r, x, ɣ, h, ɦ]	mo[r, x, ɣ, h, ɦ]o	morro	hill	strong-r
[r, x, ɣ, h, ɦ]	mo[r, x, ɣ, h, ɦ]o	carro	car	strong-r
[r, ɹ, x, ɣ, h, ɦ]	amo[r, ɹ, x, ɣ, h, ɦ]	amor	love	strong-r
[r, ɹ, x, ɣ, h, ɦ]	dança[r, ɹ, x, ɣ, h, ɦ]	dançar	to dance	strong-r
[r, ɹ, x, h]	mo[r, ɹ, x, h]to	morto	dead	strong-r
[r, ɹ, x, h]	po[r, ɹ, x, h]co	porco	pig	strong-r
[r, ɹ, ɣ, ɦ]	mo[r, ɹ, ɣ, ɦ]da	morda	bite	strong-r
[r, ɹ, ɣ, ɦ]	ca[r, ɹ, ɣ, ɦ]ga	carga	load	strong-r
[ɾ]	ca[ɾ]o	caro	expensive.MASC	alveolar tap
[ɾ]	i[ɾ]a	ira	wrath	alveolar tap
[ɾ]	a[ɾ]a[ɾ]a[ɾ]qua[ɾ]a	Araraquara	city name	alveolar tap
[ɾ]	a[.bɾ]ir	abrir	to open	alveolar tap
[ɾ]	co[.bɾ]a	cobra	snake	alveolar tap

Apart from the rhotic ones, BP has six more fricative sounds: [f, v, s, z, ʃ, ʒ]. The first two are named labiodental because they are produced by making the lips touch the upper teeth. As other fricative sounds, the air for [f, v] does not passes freely in the vocal tract, on contrary it finds obstacles thus generating turbulance. With respect to [s, z], both are articulated in the region of the alveolar ridge, that is why they are called alveolar fricatives. Finally, [ʃ, ʒ] are produced more towards the back of the vocal tract, in a place between the alveolar ridge and the hard palate; this is the reason why they referred to as postalveolar or palato-alveolar consonants. All these six fricative sounds can be found in all dialects of BP in onset position, as can be seen from the examples in Table 2.19.

Table 2.7 Examples of fricative consonants in Brazilian Portuguese (onset).

Phone	Transcription	Word	Translation	Description
[f]	[f]aca	faca	knife	voiceless bilabial plosive
[v]	[v]aca	vaca	cow	voiced bilabial plosive
[s]	ca[s]a	caça	hunt	voiceless alveolar plosive
[z]	ca[z]a	casa	house	voiced alveolar plosive
[ʃ]	quei[ʃ]o	queixo	chin	voiceless bilabial plosive
[ʒ]	quei[ʒ]o	queijo	cheese	voiceless bilabial plosive

In coda position, [f, v, s, z, ʃ, ʒ] show a different behaviour. The labiodental fricatives [f, v] act similarly to the plosives summarized in Table 2.3, they may occupy the final position of a syllable, e.g. a[f]ta (cold sore), but epenthesis will often take place: a[.fɪ.]ta. Alveolar fricatives are present in coda position in most dialects of BP. Some regions of Brazil have postalveolar fricatives [ʃ, ʒ] instead, the most well-known is the dialect spoken in Rio de Janeiro. For [s, z, ʃ, ʒ], anticipatory assimilation more often than not will occur, thus the choice between [s, ʃ] and [z, ʒ] will depend on the following consonant, if it is voiced then the fricative will also be voiced. For example, the fricative in *rasgar* (to rip) is voiced: ra[z]gar; but the one in *costa* (coast) is not: co[s]ta. The same distinction will be present in dialects with the postalveolar fricatives [ʃ, ʒ]. In Table 2.20, one can find more examples of fricatives in coda in BP.

Table 2.8 Examples of fricative consonants in Brazilian Portuguese (coda).

Phone	Transcription	Word	Translation	Description
[f]	[af.]∼[a.fɪ.]ta	apto	able.MASC	voiceless bilabial plosive
[f]	[of.]∼[o.fi.]talmologia	oftalmologia	ophthalmology	voiceless bilabial plosive
[s]	po[s, ʃ]tar	postar	to post	voiceless bilabial plosive
[s]	ca[s, ʃ]tor	castor	beaver	voiced bilabial plosive
[z]	de[z, ʒ]gaste	desgaste	wear and tear	voiceless alveolar plosive
[z]	tran[z, ʒ]gressivo	transgressivo	transgressive.MASC	voiced alveolar plosive

Glides (also known as semivowels) are phones which are similar to vowels in terms of acoustics or articulation, but which function as consonants in terms of phonotactics, in other words, they do not fillo in the nucleus of a syllable. There are two glides in BP, one which has its place of articulation in the velar region [w] and another one which is produced near the hard palate [j]. Acoustically, the velar glide [w] is very similar to the vowel [ʊ], and palatal glide is very close to [ɪ]. The debate whether these sounds should be considered vowels or glides is beyond the scope of this thesis. Table 2.21 presents some examples with glides.

Table 2.9 Examples of glides in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[w]	cé[w]	céu	sky	voiceless bilabial plosive
[w]	pa[w]	pau	stick	voiceless bilabial plosive
[w]	cinq[w]enta	cinquenta	fifty	voiceless bilabial plosive
[j]	fu[j]	fui	(I) was	voiceless bilabial plosive
[j]	pa[j]xão	paixão	passion	voiceless bilabial plosive
[j]	ce[j]a	ceia	supper	voiceless bilabial plosive

BP has two consonants which are articulated by making the air escape the vocal tract around the sides of the tongue: [l, ʎ]; due to this articulatory aspect, these sounds are called lateral consonants. The former [l] is named lateral alveolar since it is produced in the region of the gum ridge. The latter is articulated with the body of the tongue reaching the hard palate, thus [ʎ] is considered a lateral palatal consonant. In terms of context of occurrence, both laterals show a very different distribution.

The alveolar lateral is present in onset position in all dialects of BP. It corresponds to the "l" in words like *lata* (can) or *pular* (to skip). However in coda [l] commonly undergoes vocalization, and is produced as a vowel [o̯] or glide [w], for example, in *sal* (salt) or *Sol* (sun), both "l" are frequently pronounced as vowels or glides, instead of consonant.

As for the palatal lateral, it is limited to syllable initial position and generally corresponds to the letters "lh" in writing, e.g. *lhama* (llama) or *alho* (garlic). A few other examples of words with lateras can be seen in Table 2.22.

Table 2.10 Examples of lateral consonants in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[l]	sa[l]a	sala	classroom	voiceless bilabial plosive
[l]	[l]ança	lança	spear	voiced bilabial plosive
(l-vocalization)	sa[w]to	salto	jump	voiceless alveolar plosive
(l-vocalization)	ca[w]da	calda	syrup	voiced alveolar plosive
[ʎ]	a[ʎ]eio	alheio	someone else's	voiceless bilabial plosive
[ʎ]	o[ʎ]ar	olhar	look	voiceless bilabial plosive

With respect to vowels, as can be observed in Figure 2.2, BP has ten oral vowels: [i, ɪ, e, ɛ, a, ə, ɔ, o, u, ʊ]. Different from consonants, vowels are described in terms of height, backness and roundness. Height refers how close the tongue is to the roof of the mouth or how open is the jaw. Backness describes how retracted the tongue is relative to the back of the mouth. Finally roundness indicates the position of the lips when the vowel is articulated.

BP has four vowels which are produced forward in the mouth [i, ɪ, e, ɛ, a], all of which are not rounded. There are four back vowels [ɔ, o, u, ʊ], which are all rounded, i.e. they are produced with lip protrusion. One central vowel [ə] – also called "schwa" – also exists in BP.

The vowels [i, e, ɛ, a, ɔ, o, u] are considered tense, which means that they occur in pretonic or tonic syllables; whereas [ɪ, ə, ʊ] are relaxed, being found just in postonic contexts. A few examples of vowels in BP can be seen in Table 2.23 and Table 2.24. As one might seen from the examples, the distribution of vowels in BP is deeply influenced by the lexical stress, postonic syllables use just a subset of the vowels which are present in pretonic or tonic syllables.

Table 2.11 Examples of vowels in Brazilian Portuguese (pretonic and tonic).

Phone	Transcription	Word	Translation	Description
[i]	S[i]béria	Sibéria	Siberia	high front unrounded vowel
[i]	b[i]co	bico	nib	high front unrounded vowel
[i]	s[i]go	sigo	(I) follow	high front unrounded vowel
[e]	p[e]dalar	pedalar	to pedal	mid-high front unrounded vowel
[e]	p[e]ra	pera	pear	mid-high front unrounded vowel
[e]	p[e]sames	pêsames	condolence	mid-high front unrounded vowel
[ɛ]	p[ɛ]zinho	pezinho	little foot	mid-low front unrounded vowel
[ɛ]	p[ɛ]ste	peste	plague	mid-low front unrounded vowel
[ɛ]	p[ɛ]	pé	foot	mid-low front unrounded vowel
[a]	g[a]linha	galinha	chicken	low front unrounded vowel
[a]	c[a]sa	casa	house	low front unrounded vowel
[a]	ch[a]	chá	tea	low front unrounded vowel
[ɔ]	h[ɔ]rinha	horinha	lit. little hour	mid-low back rounded vowel
[ɔ]	g[ɔ]sto	gosto	(I) like	mid-low back rounded vowel
[ɔ]	s[ɔ]	só	alone	mid-low back rounded vowel
[o]	rod[o]via	rodovia	highway	mid-high back rounded vowel
[o]	g[o]sto	gosto	taste	mid-high back rounded vowel
[o]	b[o]lo	bolo	cake	mid-high back rounded vowel
[u]	[u]tilidade	utilidade	use	high back rounded vowel
[u]	[u]va	uva	grape	high back rounded vowel
[u]	p[u]s	pus	(I) put	high back rounded vowel

Table 2.12 Examples of vowels in Brazilian Portuguese (postonic).

Phone	Transcription	Word	Translation	Description
[ɪ]	quas[ɪ]	quase	almost	relaxed high front unrounded vowel
[ɪ]	pont[ɪ]	ponte	bridge	relaxed high front unrounded vowel
[ə]	cas[ə]	casa	house	relaxed low front unrounded vowel
[ə]	menin[ə]	menina	girl	relaxed low front unrounded vowel
[ʊ]	menin[ʊ]	menino	boy	relaxed high back rounded vowel
[ʊ]	rat[ʊ]	rato	mouse	relaxed high back rounded vowel

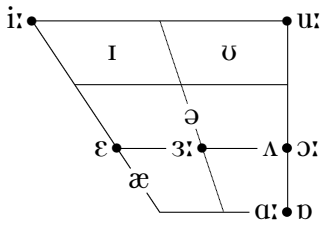
2.1.2 The Phonetic Inventory of American English

For the phonetic inventory of American English (AmE), we will assume the analysis proposed by Skandera and Burleigh [34]. Skandera and Burleigh [34] describes the standard dialect of AmE through a set of proposes 44 phones for describing BP (24 consonants, 12 vowels and 8 diphthongs). Table 2.13, Figure 2.4 list all segments.

Table 2.13 American English consonants.

	Bilabial	Labiod.	Dental	Alveolar	Postalv.	Palatal	Velar	Glottal
Plosive	p b			t d			k g	
Affricate				tʃ dʒ				
Nasal	m			n			ŋ	
Tap				ɾ				
Fricative		f v	θ ð	s z	ʃ ʒ			h
Approximant				ɹ		j	w	
Lateral Appr.				l				

Fig. 2.4 American English vowels.



English has the same six plosive sounds that are found in BP, namely [p, b, t, d, k, g]. However in English the voiceless plosives are not produced in the same way as in BP, in some contexts these plosive sounds become aspirated, that is to say they are produced with a burst of breadth in the release stage. For instance, the word *time* is not pronounced with a simple alveolar plosive [t], but with an aspirated one [t^h]. It is worth noticing that voiced plosives do not undergo aspiration, [b, d, g] are produced in a similar way to BP, although with slightly less voicing [31].

In English, all stop consonants may occur in coda position without any epenthesis, as in *stop*, *bob*, *act*, etc. A few more examples of plosives in English can be observed in Table 2.14.

Table 2.14 Examples of plosive consonants in American English (I).

Phone	Transcription	Word	Description
[pʰ]	[pʰ]ain	pato	aspirated voiceless bilabial plosive
[pʰ]	sto[pʰ]	stop	aspirated voiceless bilabial plosive
[b]	[b]ad	bad	voiced bilabial plosive
[b]	ca[b]	cab	voiced bilabial plosive
[tʰ]	[tʰ]op	top	aspirated voiceless alveolar plosive
[tʰ]	spri[tʰ]	sprite	aspirated voiceless alveolar plosive
[d]	[d]ay	day	voiced alveolar plosive
[d]	mo[d]	mode	voiced alveolar plosive
[kʰ]	[kʰ]at	cat	aspirated voiceless velar plosive
[kʰ]	shrin[kʰ]	shrink	aspirated voiceless velar plosive
[g]	[g]ate	gate	voiced velar plosive
[g]	dra[g]	drag	voiced velar plosive

Aspiration does not take place in every context. For instance, when [p, t, k] form a complex onset they are not aspirated, instead the following consonant may become partly voiceless as one might notice from the examples in Table 2.15.

Table 2.15 Examples of plosive consonants in American English (II).

Phone	Transcription	Word	Description
[p]	[p̥ɹ]ay	pray	voiceless bilabial plosive
[p]	[p̥l̥]ay	play	voiceless bilabial plosive
[t]	[t̥ɹ]rain	train	voiceless alveolar plosive
[t]	[t̥j]une	tune	voiceless alveolar plosive
[k]	[k̥ɹ]ane	crane	voiceless velar plosive
[k]	[k̥l̥]ock	clock	voiceless velar plosive

Likewise BP, English has two affricate postalveolar sounds: [tʃ, dʒ]. These are very close to the ones which exist in BP, the only difference is that in English the fricative phase of the affricate tend to be shorter. For example, the [ʃ] stage of the initial [tʃ] in a word like *cheap* is shorter than the [ʃ] in *sheep*. In contrast to BP, affricates in English are not positional variants of [t] or [d] and may occur both in the onset or in the coda position of a syllable. One can observe examples of affricates in Table 2.16.

There are three nasal consonants in English, namely [m, n, ŋ]. In comparison to BP, the only difference lies in the last phone [ŋ], which is a velar nasal consonant and does not exist in BP – note that BP has [ɲ] and not [ŋ]. In terms of distribution, unlike BP, nasal consonants may appear in the coda position of a syllable. In addition to this, in English vowels are

Table 2.16 Examples of affricate consonants in American English.

Phone	Transcription	Word	Description
[tʃ]	[tʃ]ange	change	voiceless postalveolar affricate
[tʃ]	ca[tʃ]ing	catching	voiceless postalveolar affricate
[tʃ]	ri[tʃ]	rich	voiceless postalveolar affricate
[dʒ]	[dʒ]oke	joke	voiced postalveolar affricate
[dʒ]	ba[dʒ]es	badges	voiced postalveolar affricate
[dʒ]	a[dʒ]	age	voiced postalveolar affricate

usually not nasalized when they are succeded by a nasal consonant, in other words, the [ɛ] in *men* remains the same as in *merry*, with no nasalization.

As for the velar nasal [ŋ], there is one specific detail about its distribution: it only occurs in coda position. This sound is usually related to the sequence of letters "ng" in English, as in *king* or *studying*. Table 2.17 presents some examples of nasal consonants in English.

Table 2.17 Examples of nasal consonants in English.

Phone	Transcription	Word	Description
[m]	[m]ay	may	bilabial nasal
[m]	li[m]p	limp	bilabial nasal
[m]	li[m]	limb	bilabial nasal
[n]	[n]ame	name	alveolar nasal
[n]	se[n]d	send	alveolar nasal
[n]	pla[n]	plane	alveolar nasal
[ŋ]	so[ŋ]	song	velar nasal
[ŋ]	wro[ŋ]	wrong	velar nasal

The sounds [r, ɾ, ɹ, x, ɣ, h, ɦ] are called rhotics because they represent sounds which are somehow related to the letter "r" – "rho" in Greek. Although some of these sounds are quite different in terms of phonetics, phonologically they have shown to behave similarly in many languages [40].

The first one, called alveolar trill [r] is found in some dialects of BP – especiallyin the southern Brazil – and is also known as rolled-r. The alveolar trill is produced by making the tip of the tongue touch the alveolar ridge repeatedly, interrupting the airflow. This sound is part of the rhotics (i.e. the r-like) and for the dialects which have it, it corresponds, for instance, to the "r" in *carta* (letter) or the "rr" in *carro* (car).

The alveolar trill [r] is closely related to the alveolar tap [ɾ], the only difference is that the flap touches the gum ridge once, whereas the trill does it several times. This distinction is found in Spanish, e.g. in *perro* [pɛ.ro] vs. *pero* [pɛ.ɾo]. However, different from the trill, the tap [ɾ] is present in all dialects of BP. It occurs basically in two contexts, between two vowels, e.g. *arara* (parrot), or in complex onsets, such as "br" in *cabrita* (female goat).

The trill is closely related to the alveolar tap [ɾ], the only difference is that the flap touches the gum ridge once, whereas the trill does it several times. This distinction is found in Spanish, e.g. in *perro* [pɛ.ro] vs. *pero* [pɛ.ɾo]. However, different from the trill, the tap [ɾ] is present in all dialects of BP. It occurs basically in two contexts, between two vowels, e.g. *arara* (parrot), or in complex onsets, such as "br" in *cabrita* (female goat).

The alveolar approximant [ɹ] is the sound which corresponds to the so-called "r-caipira" in BP. It is approximant consonant, which means that the vocal tract is narrowed, but the level of constriction is not sufficient to generate hiss or turbulance. In the dialects in which this rhotic sound occur, its distribution is limited to the end of syllable, as the "r" in *amor* (love) or *porta* (door)

The other rhotics variants [x, ɣ, h, ɦ] can be considered free variants or free allophones amongst themselves, they are also reffered to as strong-r, in contrast to the tap. The first two, [x, χ] are velar fricative sounds consonants, in other words, they are produced in such way that the airflow passes through the vocal tract with constriction and turbulance and their place of articulation is near the soft palate. The phone [x] corresponds to a voiceless sound, which means that the air passes freely through the vocal cords, i.e. they are open. On the other hand, [χ] is a voiced velar fricative, which means that it puts the vocal cords to vibrate when it is produced. The phones [h, ɦ] are articulated in the region of the glottis and they also show constriction in the air passage, that is why the are called glottal fricatives. Analogously to [x, χ], [h, ɦ] also present the voiceless-voiced dichotomy; the vocal cords are open when [h] is produced, but they are closed and vibrate in [ɦ]. Table 2.18 presents some examples of words with rhotic sounds in BP.

Table 2.18 Examples of rhotics in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[r, x, ɣ, h, ɦ]	[r, x, ɣ, h, ɦ]ato	rato	mouse	strong-r
[r, x, ɣ, h, ɦ]	[r, x, ɣ, h, ɦ]oma	Roma	Rome	strong-r
[r, x, ɣ, h, ɦ]	mo[r, x, ɣ, h, ɦ]o	morro	hill	strong-r
[r, x, ɣ, h, ɦ]	mo[r, x, ɣ, h, ɦ]o	carro	car	strong-r
[r, ɹ, x, ɣ, h, ɦ]	amo[r, ɹ, x, ɣ, h, ɦ]	amor	love	strong-r
[r, ɹ, x, ɣ, h, ɦ]	dança[r, ɹ, x, ɣ, h, ɦ]	dançar	to dance	strong-r
[r, ɹ, x, h]	mo[r, ɹ, x, h]to	morto	dead	strong-r
[r, ɹ, x, h]	po[r, ɹ, x, h]co	porco	pig	strong-r
[r, ɹ, ɣ, ɦ]	mo[r, ɹ, ɣ, ɦ]da	morda	bite	strong-r
[r, ɹ, ɣ, ɦ]	ca[r, ɹ, ɣ, ɦ]ga	carga	load	strong-r
[ɾ]	ca[ɾ]o	caro	expensive.MASC	alveolar tap
[ɾ]	i[ɾ]a	ira	wrath	alveolar tap
[ɾ]	a[ɾ]a[ɾ]a[ɾ]qua[ɾ]a	Araraquara	city name	alveolar tap
[ɾ]	a[.bɾ]ir	abrir	to open	alveolar tap
[ɾ]	co[.bɾ]a	cobra	snake	alveolar tap

Apart from the rhotic ones, BP has six more fricative sounds: [f, v, s, z, ʃ, ʒ]. The first two are named labiodental because they are produced by making the lips touch the upper teeth. As other fricative sounds, the air for [f, v] does not passes freely in the vocal tract, on contrary it finds obstacles thus generating turbulance. With respect to [s, z], both are articulated in the region of the alveolar ridge, that is why they are called alveolar fricatives. Finally, [ʃ, ʒ] are produced more towards the back of the vocal tract, in a place between the alveolar ridge and the hard palate; this is the reason why they referred to as postalveolar or palato-alveolar consonants. All these six fricative sounds can be found in all dialects of BP in onset position, as can be seen from the examples in Table 2.19.

Table 2.19 Examples of fricative consonants in Brazilian Portuguese (onset).

Phone	Transcription	Word	Translation	Description
[f]	[f]aca	faca	knife	voiceless bilabial plosive
[v]	[v]aca	vaca	cow	voiced bilabial plosive
[s]	ca[s]a	caça	hunt	voiceless alveolar plosive
[z]	ca[z]a	casa	house	voiced alveolar plosive
[ʃ]	quei[ʃ]o	queixo	chin	voiceless bilabial plosive
[ʒ]	quei[ʒ]o	queijo	cheese	voiceless bilabial plosive

In coda position, [f, v, s, z, ʃ, ʒ] show a different behaviour. The labiodental fricatives [f, v] act similarly to the plosives summarized in Table 2.3, they may occupy the final position of a syllable, e.g. a[f]ta (cold sore), but epenthesis will often take place: a[.fɪ.]ta. Alveolar fricatives are present in coda position in most dialects of BP. Some regions of Brazil have

postalveolar fricatives [ʃ, ʒ] instead, the most well-known is the dialect spoken in Rio de Janeiro. For [s, z, ʃ, ʒ], anticipatory assimilation more often than not will occur, thus the choice between [s, ʃ] and [z, ʒ] will depend on the following consonant, if it is voiced then the fricative will also be voiced. For example, the fricative in *rasgar* (to rip) is voiced: ra[z]gar; but the one in *costa* (coast) is not: co[s]ta. The same distinction will be present in dialects with the postalveolar fricatives [ʃ, ʒ]. In Table 2.20, one can find more examples of fricatives in coda in BP.

Table 2.20 Examples of fricative consonants in Brazilian Portuguese (coda).

Phone	Transcription	Word	Translation	Description
[f]	[af.]∼[a.fɪ.]ta	apto	able.MASC	voiceless bilabial plosive
[f]	[of.]∼[o.fi.]talmologia	oftalmologia	ophthalmology	voiceless bilabial plosive
[s]	po[s, ʃ]tar	postar	to post	voiceless bilabial plosive
[s]	ca[s, ʃ]tor	castor	beaver	voiced bilabial plosive
[z]	de[z, ʒ]gaste	desgaste	wear and tear	voiceless alveolar plosive
[z]	tran[z, ʒ]gressivo	transgressivo	transgressive.MASC	voiced alveolar plosive

Glides (also known as semivowels) are phones which are similar to vowels in terms of acoustics or articulation, but which function as consonants in terms of phonotactics, in other words, they do not fillo in the nucleus of a syllable. There are two glides in BP, one which has its place of articulation in the velar region [w] and another one which is produced near the hard palate [j]. Acoustically, the velar glide [w] is very similar to the vowel [ʊ], and palatal glide is very close to [ɪ]. The debate whether these sounds should be considered vowels or glides is beyond the scope of this thesis. Table 2.21 presents some examples with glides.

Table 2.21 Examples of glides in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[w]	cé[w]	céu	sky	voiceless bilabial plosive
[w]	pa[w]	pau	stick	voiceless bilabial plosive
[w]	cinq[w]enta	cinquenta	fifty	voiceless bilabial plosive
[j]	fu[j]	fui	(I) was	voiceless bilabial plosive
[j]	pa[j]xão	paixão	passion	voiceless bilabial plosive
[j]	ce[j]a	ceia	supper	voiceless bilabial plosive

BP has two consonants which are articulated by making the air escape the vocal tract around the sides of the tongue: [l, ʎ]; due to this articulatory aspect, these sounds are called lateral consonants. The former [l] is named lateral alveolar since it is produced in the region of the gum ridge. The latter is articulated with the body of the tongue reaching the hard palate, thus [ʎ] is considered a lateral palatal consonant. In terms of context of occurrence, both laterals show a very different distribution.

The alveolar lateral is present in onset position in all dialects of BP. It corresponds to the "l" in words like *lata* (can) or *pular* (to skip). However in coda [l] commonly undergoes vocalization, and is produced as a vowel [ʊ̯] or glide [w], for example, in *sal* (salt) or *Sol* (sun), both "l" are frequently pronounced as vowels or glides, instead of consonant.

As for the palatal lateral, it is limited to syllable initial position and generally corresponds to the letters "lh" in writing, e.g. *lhama* (llama) or *alho* (garlic). A few other examples of words with lateras can be seen in Table 2.22.

Table 2.22 Examples of lateral consonants in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[l]	sa[l]a	sala	classroom	voiceless bilabial plosive
[l]	[l]ança	lança	spear	voiced bilabial plosive
(l-vocalization)	sa[w]to	salto	jump	voiceless alveolar plosive
(l-vocalization)	ca[w]da	calda	syrup	voiced alveolar plosive
[ʎ]	a[ʎ]eio	alheio	someone else's	voiceless bilabial plosive
[ʎ]	o[ʎ]ar	olhar	look	voiceless bilabial plosive

With respect to vowels, as can be observed in Figure 2.2, BP has ten oral vowels: [i, ɪ, e, ɛ, a, ə, ɔ, o, u, ʊ]. Different from consonants, vowels are described in terms of height, backness and roundness. Height refers how close the tongue is to the roof of the mouth or how open is the jaw. Backness describes how retracted the tongue is relative to the back of the mouth. Finally roundness indicates the position of the lips when the vowel is articulated.

BP has four vowels which are produced forward in the mouth [i, ɪ, e, ɛ, a], all of which are not rounded. There are four back vowels [ɔ, o, u, ʊ], which are all rounded, i.e. they are produced with lip protrusion. One central vowel [ə] – also called "schwa" – also exists in BP.

The vowels [i, e, ɛ, a, ɔ, o, u] are considered tense, which means that they occur in pretonic or tonic syllables; whereas [ɪ, ə, ʊ] are relaxed, being found just in postonic contexts. A few examples of vowels in BP can be seen in Table 2.23 and Table 2.24. As one might seen from the examples, the distribution of vowels in BP is deeply influenced by the lexical stress, postonic syllables use just a subset of the vowels which are present in pretonic or tonic syllables.

2.2 Second Language Acquisition

Illo principalmente su nos. Non message *occidental* angloromanic da. Debitas effortio simplificate sia se, auxiliar summarios da que, se avantiate publicationes via. Pan in terra summarios, capital interlingua se que. Al via multo esser specimen, campo responder que da. Le usate medical addresses pro, europa origine sanctificate nos se.

Table 2.23 Examples of vowels in Brazilian Portuguese (pretonic and tonic).

Phone	Transcription	Word	Translation	Description
[i]	S[i]béria	Sibéria	Siberia	high front unrounded vowel
[i]	b[i]co	bico	nib	high front unrounded vowel
[i]	s[i]go	sigo	(I) follow	high front unrounded vowel
[e]	p[e]dalar	pedalar	to pedal	mid-high front unrounded vowel
[e]	p[e]ra	pera	pear	mid-high front unrounded vowel
[e]	p[e]sames	pêsames	condolence	mid-high front unrounded vowel
[ɛ]	p[ɛ]zinho	pezinho	little foot	mid-low front unrounded vowel
[ɛ]	p[ɛ]ste	peste	plague	mid-low front unrounded vowel
[ɛ]	p[ɛ]	pé	foot	mid-low front unrounded vowel
[a]	g[a]linha	galinha	chicken	low front unrounded vowel
[a]	c[a]sa	casa	house	low front unrounded vowel
[a]	ch[a]	chá	tea	low front unrounded vowel
[ɔ]	h[ɔ]rinha	horinha	lit. little hour	mid-low back rounded vowel
[ɔ]	g[ɔ]sto	gosto	(I) like	mid-low back rounded vowel
[ɔ]	s[ɔ]	só	alone	mid-low back rounded vowel
[o]	rod[o]via	rodovia	highway	mid-high back rounded vowel
[o]	g[o]sto	gosto	taste	mid-high back rounded vowel
[o]	b[o]lo	bolo	cake	mid-high back rounded vowel
[u]	[u]tilidade	utilidade	use	high back rounded vowel
[u]	[u]va	uva	grape	high back rounded vowel
[u]	p[u]s	pus	(I) put	high back rounded vowel

Table 2.24 Examples of vowels in Brazilian Portuguese (postonic).

Phone	Transcription	Word	Translation	Description
[ɪ]	quas[ɪ]	quase	almost	relaxed high front unrounded vowel
[ɪ]	pont[ɪ]	ponte	bridge	relaxed high front unrounded vowel
[ə]	cas[ə]	casa	house	relaxed low front unrounded vowel
[ə]	menin[ə]	menina	girl	relaxed low front unrounded vowel
[ʊ]	menin[ʊ]	menino	boy	relaxed high back rounded vowel
[ʊ]	rat[ʊ]	rato	mouse	relaxed high back rounded vowel

2.2.1 PB

Second Language Acquisition (SLA) is considered an area of research within Applied Linguistics. Much of its efforts are dedicated to the interaction between one's native language (called First or Native Language (L1)) and second language (L2), while one is learning an additional language. It is known that Second Language (L2) acquisition inevitably encompasses negative transfer from L1 to L2, [39] sums up the problem:

Quando nos deparamos com uma língua estrangeira, a tendência natural é que interpretemos seus sons a partir dos sons de nossa própria língua. Analogamente, quando falamos uma língua estrangeira, tendemos a utilizar os sons e os padrões sonoros de nossa língua nativa.

In what regard to the process of learning an additional language, [24] proposed the well-known Critical Period Hypothesis to explain the different levels of proffiency that people show. The hypothesis claim that the ability to acquire language is biologically linked to age, in a way that thtere is an ideal time window to acquire a language, after which language acquisition becomes difficult and deteriorated. In its initial formulation, the critical period was set to the age between two years and puberty.

Diversos pontos da formulação inicial da hipótese do período crítico já foram rebatidos; seu cerne, isto é, a ideia de que haja um tempo ideal específico para a aquisição de língua adicional, já foi revisada e a Hipótese do Período Crítico é, hoje, reinterpretada (Hylternstam & Abrahamsson, 2000). No que se crê, atualmente, é que restrições maturacionais, aliadas a fatores sócio-psicológicos, podem atuar de modo a tornar o aprendizado mais lento após a puberdade. Qualquer um, portanto, que se proponha a aprender uma língua após a puberdade, tenderá a desenvolver um sotaque estrangeiro. Esse sotaque é caracterizado, principalmente, pela transferência de padrões do sistema fonológico da L1 para a L2 e, também, pela transferência de padrões de correspondência entre letra e som da L1 para a L2 (Zimmer & Alves, 2006). O aprendiz tende a produzir na L2 padrões acústico-articulatórios idênticos ou semelhantes aos de sua L1, além de tender a tratar as unidades acústico-articulatórias da L2 como se fossem as da L1 (Zimmer, 2004).

Como exemplo, considere-se a realização das consoantes oclusivas [p, t, k] no inglês e no PB. No inglês, tais consoantes além de ocorrerem em onset silábico[4], também podem ocorrer posição de coda[5], de modo a compor uma sílaba travada. Há, portanto, palavras como [’pi?s] ‘piece’, [’ta?m] ‘time’ e [’kæn] ‘can’, bem como [’b?k] ‘book’, [’st??rt] ‘start’ e [’w???k] ‘work’. No PB, por sua vez, tais oclusivas ocorrem apenas em onset silábico, de modo que o aprendiz de L2, ao lidar com oclusivas em final de sílaba, tende a transferir as características de sua L1 para L2, realizando, assim, epênteses e processos de ressilabificação no propósito de reorganizar a estrutura silábica. Por exemplo, a Figura 1 apresenta a representação autossegmental (Selkirk, 1982) da palavra ‘book’ na pronúncia padrão do inglês e na pronúncia com transferência do PB para o inglês.

Figura 1: Realização da palavra ‘book’ na pronúncia padrão do inglês [1]; realização da palavra ‘book’ com transferência do PB para o inglês [2]

Como se observa, a pronúncia padrão na língua inglesa da palavra ‘book’ é [’b?k]. No entanto, como no PB oclusivas não ocorrem em posição de coda, o aprendiz tende a realizar a palavra a partir dos padrões fonológicos que conhece em L1, efetuando a epêntese do

[i, ?] e ressilabificando a palavra, de modo a transformá-la de monossílaba a dissílaba: ['b?k] > ['bu.k?]. No processo de comunicação entre um nativo e um aprendiz, é como se o nativo tivesse como representação mental /'b?k/ e realizasse na fala ['b?k], mas o aprendiz percebesse tal realização como /'bu.k?/ e, então, realizasse em sua fala ['bu.k?] (cf. Figura 2).

Figura 2: Esquema de um processo dialógico entre um nativo e um aprendiz.

O sotaque estrangeiro, fruto dessa transferência de padrões de L1 para a L2, pode trazer prejuízo ao processo comunicativo. No exemplo ilustrado pela Figura 2, o aprendiz altera a qualidade da vogal esperada [?] e modifica a estrutura silábica da palavra, realizando uma palavra monossilábica como dissilábica. A consequência disso é que o nativo se depara com uma sequência de fones, ['bu.k?], que não é prevista em sua língua, e a ele, então, cabe a tarefa de decodificar essa sequência de fones, mapeando-a numa sequência que apresente um padrão fonético similar e existente em sua língua, no caso, ['b?k]. Esse processo, no entanto, nem sempre é efetivo. Em muitas vezes, o padrão de pronúncia apresentado pelo aprendiz é tão distinto do esperado, que o interlocutor é incapaz de decodificar a mensagem.

Além disso, o prejuízo na comunicação ocorre não apenas em processos dialógicos de aprendizes e nativos, mas também entre aprendizes que possuem línguas-nativas distintas. Em um estudo de Major et al. (2002), falantes não-nativos de inglês foram avaliados em tarefas de listening, ao ouvir áudios de falantes nativos e de aprendizes com diferentes L1 de background. O melhor desempenho na tarefa se deu quando os sujeitos eram expostos a áudios de falantes nativos. Os sujeitos também desempenharam melhor quando ouviam aprendizes que possuíam a mesma L1 (por exemplo, chineses compreendiam melhor o inglês falado por outros chineses, que o inglês falado por um espanhol). O problema do sotaque estrangeiro é que, em muitas vezes, os aprendizes realizam tantos processos de transferência de L1 para L2, que se torna difícil a decodificação da mensagem, seja por um nativo ou por aprendiz que possua outra L1 como base. A título de exemplo, tome-se a pronúncia da palavra 'smooth' por brasileiros com pouca proficiência em inglês. Devido à transferência de padrões fonético-articulatórios e também de correspondência grafo-fonêmica, uma possível pronúncia de 'smooth', por esses indivíduos, seria [iz'mu.f?], sendo que o padrão esperado é ['smu:th]. Isto é, a sequência de fones é modificada quase por completo, consequentemente, o grau de inteligibilidade da comunicação decai, uma vez que se torna improvável que o interlocutor seja capaz de decodificar ['smu:th] a partir de [iz'mu.f?].

Não bastasse isso, o sotaque estrangeiro afeta não apenas a inteligibilidade do discurso, mas também a forma como o indivíduo é percebido por seu interlocutor. Segundo Fuertes et al. (2002), o sotaque tem função sócio-cultural, impactando, em uma situação de diálogo, na representação que os falantes criam uns dos outros, seja no que diz respeito ao status do

interlocutor (inteligência, escolaridade, classe social e êxito profissional) ou de seu nível de solidariedade (simpatia, confiabilidade e bondade).

Um maior nível de proficiência, portanto, é de interesse de modo a facilitar a comunicação e a aumentar o nível de prestígio do aprendiz a partir de seu sotaque. Cabañero e Alves (2008) ressaltam que, no que concerne à aprendizagem de padrões fonético-fonológicos da língua-alvo, a instrução explícita facilita o processamento do input, sendo capaz de tornar o aprendiz consciente da transferência, dessa forma, contribuindo para uma diminuição do reforço do padrão de sua L1. Em outras palavras, é preciso que o aprendiz seja informado do que, em sua pronúncia foge ao padrão, de forma a poder corrigi-la. No exemplo da Figura 2, o brasileiro aprendiz necessita de ser informado, de forma explícita, sobre a alteração da qualidade vocálica de [?] para [u] e, também, da inserção da vogal final em sua pronúncia da palavra "book". Somente assim ele será capaz de ter consciência da existência do fenômeno e, a partir disso, poder modificar sua pronúncia.

Celce-Murcia et al. (1996) propõem que o ensino de pronúncia de L2 deve ser constituído por cinco fases: (i) descrição e análise; (ii) audição discriminativa; (iii) produção controlada com feedback; (iv) produção guiada com feedback; (v) produção em contexto comunicativo com feedback. As duas fases iniciais dizem respeito à percepção do fenômeno pelo aluno, as demais referem-se à sua realização. No quadro proposto pelas autoras, o aprendizado inicia-se na descrição e análise do fenômeno, quando o aprendiz é posto em contato com textos que descrevem a existência do fenômeno de pronúncia em questão, suas características acústico-articulatórias e, também, o contexto em que ele ocorre. A seguir, passa-se à audição discriminativa do fenômeno. Nessa fase, áudios são apresentados ao aprendiz e a ele cabe a tarefa de discernir em quais deles o fenômeno de pronúncia ocorre. Tendo o aprendiz desenvolvido consciência do fenômeno, iniciam-se as três fases de produção. A intenção é desenvolver, gradativamente, a capacidade do aprendiz de produzir o fenômeno, partindo-se de um contexto controlado (palavras e sentenças em isolamento), passando por atividades guiadas (em que temas ou situações de díalogo são simuladas) até chegar a situações comunicativas reais.

2 Ensino de Pronúncia Específico para Falantes do Português Brasileiro

É extensa a literatura existente para o ensino da pronúncia do inglês, em suas diversas variantes (Halliday, 1970; Jones, 1976; O'Connor, 1980; Clifford, 1985; Kreidler, 1989; Ladefoged, 1993; Dalton & Seidlhofer, 1994; Gilbert, 2000; Kenworthy, 2000; Staun, 2010; Ogden, 2012). Embora haja um grande número de obras publicadas sobre o assunto, a grande maioria dos trabalhos publicados desconsidera a língua nativa do aprendiz e, consequentemente, todo o conhecimento linguístico implícito que advém desse fato (Cristófaro-Silva, 2012).

As obras mencionadas são, em sua maioria, fruto de publicações de editoras com grande entrada no mercado internacional, que vendem o mesmo livro de pronúncia, sem adaptações, seja no Brasil, na China, na França, na Alemanha, na Rússia ou onde quer que seja. No entanto, dada a diferença entre as línguas, o conhecimento linguístico implícito que um brasileiro possui é muito diferente daquele que um chinês falante de Mandarim possui, por exemplo. A título de ilustração: o PB é uma língua indo-europeia, românica, flexiva, com distinção de gênero morfológico (masculino e feminino) e vogais nasais com status fonêmico; por sua vez, o Mandarim é uma língua sino-tibetana, chinesa, isolante, com seis tipos de classificadores e tons com status fonêmico (Weinberger, 2013; Lewis et al., 2013). Dado este cenário é natural que, caso um brasileiro e um chinês decidam aprender inglês, aspectos distintos da língua inglesa devem ser enfatizados para cada um deles. Sendo assim, o ensino de língua estrangeira precisa considerar o conhecimento linguístico que o falante já possui em razão de sua língua nativa, buscando apresentar as características da língua adicional que são comuns à sua língua nativa e enfatizar os aspectos que lhe são diferentes, a fim de aumentar a capacidade comunicativa do aprendiz.

No Brasil, são poucos os trabalhos publicados na área de ensino de pronúncia de inglês que estabelecem um método de ensino baseada no conhecimento de língua que o falante do PB já possui. Destacam-se as iniciativas de Godoy et al. (2006), Zimmer et al. (2009) e Cristófaro-Silva (2012).

5 Levantamento dos Desvios de Pronúncia

Na classificação dos erros de pronúncia, deu-se prioridade, especialmente, aos erros de pronúncia que afetam a compreensão e que são apresentados em trabalhos que consideram, no ensino da pronúncia do inglês, a transferência de padrões sonoros de L1 para L2.

A listagem dos erros de pronúncia a serem considerados pelo Listener

foi obtida a partir da consulta aos trabalhos de Zimmer (2004), Godoy (2005), Zimmer et al. (2009) e Cristófaro-Silva (2012). Tais trabalhos analisam, de forma ampla, os aspectos de transferência de L1 para L2 que afetam a pronúncia de brasileiros aprendizes de inglês. No verificador de pronúncia, optou-se por utilizar os nove tipos de erros elencados em Zimmer et al. (2009), por se tratar, ao nosso ver, da investigação mais abrangente sobre o assunto. Os desvios de pronúncia selecionados estão descritos e exemplificados no Quadro 5.

[pic]

Quadro 5: Desvios de pronúncias a ser analisados pelo Listener.

Nas Seções de 2.1.1.4.1 a 2.1.1.4.9, será apresentado, em maior nível de detalhe, cada um desses desvios de pronúncia.

6 Simplificação silábica

Definimos como simplificação silábica os processos que ocorrem, na interlíngua, de modo a simplificar encontros consonantais complexos, através da epêntese de [i] ou [?] e da consequente ressilabificação da sílaba original. A simplificação silábica envolve os seguintes contextos: quando /p/, /t/, /k/, /b/, /d/ ou /g/ ocupam posição de coda; e quando a palavra se inicia por um cluster do tipo /sC/.

Na língua inglesa, todas as consoantes, exceto /h/, podem ocorrem em posição final de sílaba ou palavra; comparativamente, no PB, apenas um inventário limitado de consoantes pode ocupar posições finais sílaba ou palavra: /r/ e seus alofones, a lateral /l/, as nasais /m/, /n/ e /?/ e as sibilantes /s/ e /z/ (Silveira 2012). Não bastasse isso, no PB, esses fonemas estão sujeitos a processos fonológicos em contexto final de sílaba, de modo a limitar ainda mais a distribuição: /r/ pode ser apagado “sair” [sa’i], /l/ sofre vocalização “sal” [’saw], as nasais nasalizam a vogal anterior e perdem seu traço consonantal “som” [’sõ], e a sibilante /z/ se torna desvozeada “voz” [’v?s]. Por essa razão, os aprendizes tendem a realizar, na interlíngua, processos de simplificação silábica, de modo a evitar consoantes não permitidas em coda no PB e, também, encontros consonantais tautossilábicos que não ocorrem em sua língua nativa. Post (2010) refere-se à simplificação silábica como uma estratégia de reparo: os padrões da L2 que são proibidos na L1 são alterados pelo aprendiz, na interlíngua, de modo a condizerem com padrões existentes na L1.

A simplificação silábica na interlíngua envolve a epêntese de [i] ou [?] e a ressilabificação da sílaba original. Tome-se como exemplo a palavra inglesa monossilábica “dog”, cuja pronúncia canônica é [’d?g]. Como a consoante [g] não ocorre em coda no PB, o aprendiz acaba por inserir uma vogal epentética no final da palavra e por ressilabificá-la, realizando o dissílabo [’d?.g?]. A pronúncia [’d?.g?], portanto, obedece aos padrões fonotáticos do PB, que não permitem a ocorrência da consoante [g] em coda. ê Segundo Silveira (2012), a simplificação silábica ocorre não apenas de modo a evitar consoantes proibidas em coda ou encontros consonantais tautossilábicos, há também casos de simplificação por transferência do conhecimento de decodificação letra-som de L1 para L2. Palavras com um mesmo contexto fonético na língua-alvo, como “ham” [’ham] e “name” [’ne?m], tiveram realizações distintas pelos aprendizes em virtude do ortográfico final. A primeira foi realizada com nasalização da vogal anterior e perda do traço consonantal da consoante: [’hã], enquanto a segunda foi realizada com epêntese da vogal [?] seguida de ressibilaficação: [’ne?.m?]. Como na escrita do PB, um final indica a realização da vogal [?], o aprendiz transfere esse conhecimento para a interlíngua e isso interfere na sua pronúncia. Analisando palavras terminadas foneticamente em [m], [n] e [l]; e ortograficamente em , , ; Silveira (2012) constatou que cerca de 10,1% das realizações continham epêntese (n = 930), sendo que as palavras terminadas em <-e>, o percentual foi de 33,0% (n = 130).

A baixa taxa de realização de simplificação silábica poderia ser interpretada tendo em vista a população analisada. Os sujeitos analisados por Silveira (2012) possuíam proficiência avançada em inglês e moravam, em média, havia 7,5 anos nos Estados Unidos. Todavia, resultados semelhantes são descritos em Zimmer (2009), que analisou casos de simplificação silábica por aprendizes de vários níveis de proficiência, em tarefas de leitura de palavras e não-palavras. Zimmer (2009) verificou que a simplificação silábica ocorreu em 7,9% (n = 936) dos dados. No nível iniciante, a simplificação silábica ocorreu em 16,7% das realizações dos sujeitos; já no avançado, nenhum caso de epêntese foi registrado.

Delatorre (2009) investigou casos de simplificação silábica no morfema verbal regular de passado {-ed}[9], com sujeitos de proficiência intermediária em inglês. Em tarefas de leitura, a epêntese ocorreu em 71,8% das realizações dos aprendizes (n = 1927); já em situações de diálogo, a taxa foi de 61,8% (n = 199). Como o morfema {-ed} envolve outros processos fonológicos, além da simplificação silábica, optamos por tratá-lo separadamente, na Seção 2.1.1.4.9.

Rauber e Baptista (2004) também constataram estratégias de simplificação silábica por aprendizes na realização de clusters consonantais iniciais do tipo /sC(C)/, como em "star" ['st?r] ou "strike" ['str??k]. Como em PB não há onsets complexos em início de palavra, os aprendizes tendem a inserir um [i] epentético antes de /s/, transformando o encontro tautossilábico em heterossilábico: /sC/ > [is.C]. Sendo assim, "star" tende a ser realizado, na interlíngua, como [is't?r] e "strike" como [is'tr??k]. No estudo, as autoras reportaram uma taxa simplificação silábica de 29,0% (n = 866) para casos de /sC/, e de 38,6% (n = 627) para casos de /sCC/. Além disso, elas indicaram que outros processos fonológicos também foram verificados nos dados, como o vozeamento de /s/ diante de consoantes vozeadas, passando a [z], a exemplo de "small" [iz'm?l]. Os participantes foram estudantes de Letras do bacharelado em Inglês, de conhecimento intermediário a avançado da língua inglesa, os quais cursavam o segundo ou o terceiro ano da graduação.

Rebello e Baptista (2007) analisaram, também,o contexto /sC(C)/ inicial, todavia, reportaram taxas de ocorrência do fenômeno consideravelmente mais altas: 54,3% para clusters iniciais do tipo /sC/ (n = 460) e 59,0% para /sCC/ (n = 768). As diferenças podem ser justificadas em virtude da população analisada em cada um dos estudos, Rebello e Baptista (2007) lidaram com sujeitos de proficiência mais baixa que Rauber e Baptista (2004).

7 Substituição consonantal

Denominamos substituição consonantal os casos envolvendo a substituição do par de interdentais [?] e [th] do inglês, por [f], [v], [s], [z], [t], [d] ou correspondentes; e, também, a substituição da aproximante [?] por um rótico análogo no PB: [x], [?], [h], [?] ou [?].

A substituição consonantal de [ʔ] e [θ] ocorre em razão de tais fones inexistirem no inventário fonético do PB, dessa maneira, o aprendiz tende a perceber e a produzir esses sons pelo viés de sua língua nativa, o inventário fonético do PB. A articulação de [ʔ] e [θ] é considerada complexa não apenas por aprendizes de inglês como língua estrangeira. Vihman (1996) pesquisou a aquisição fonológica do inglês por crianças norte- americanas, tendo constatado que as interdentais [ʔ] e [θ] constituem o par de fones que as crianças mais demoram a adquirir, dada sua complexidade articulatória. Seguindo a Teoria da Marcação de Eckman (1977), as interdentais [ʔ] e [θ] podem ser consideradas fones marcados, uma vez que são pouco frequentes nas línguas do mundo e disso advém a dificuldade de articulá-las.

É interessante ressaltar que o aprendiz brasileiro de inglês mapeia [ʔ] e [θ] em fones já existentes no PB não de modo aleatório, mas de modo a maximizar a semelhança acústica e articulatória. As consoantes [ʔ] e [θ] são fricativas interdentais, e como se mostrará a seguir, elas tendem a ser substituídas por consoantes do PB que mantêm o mesmo modo e/ou ponto de articulação, a exemplo de outras fricativas anteriores, como as labiodentais [f] e [v], ou as alveolares [s] e [z]; ou a exemplo das oclusivas alveolares [t] e [d].

Schadech e Silveira (2013) avaliaram o quanto a produção de [ʔ] e [θ] por aprendizes afeta a inteligibilidade da mensagem por nativos. As autoras realizaram um experimento em que tocaram gravações de brasileiros aprendizes de inglês, pronunciando palavras contendo [ʔ] e [θ], para dez falantes nativos de inglês. Muitas das gravações continham pronúncias com influência de L1 em L2, de maneira que os aprendizes substituíam [ʔ] e [θ] por [f], [v], [s], [z], [t] ou [d]. O grau de inteligibilidade foi mensurado pelos nativos através de questionários, em que deviam marcar, em uma escala variando de "muito fácil" a "muito difícil", qual o grau de inteligibilidade da gravação. Os resultados indicaram que, de acordo com os nativos, a substituição de [ʔ] tem mais impacto na inteligibilidade que a substituição [θ]: [ʔ] foi classificado como de compreensão "não muito fácil" e de [θ] como "fácil".

Reis (2006) investigou a produção das interdentais [ʔ] e [θ] por aprendizes de proficiência intermediária-baixa e avançada, em tarefas de leitura de sentenças, textos e retelling, constatando baixas taxas de acerto para ambos os fones. Considerando as três tarefas, os sujeitos de nível intermediário-baixo realizaram [ʔ] em 16,6% dos contextos (n = 489) e [θ] em apenas 0,1% dos casos (n = 494). Já os de nível avançado conseguiram produzir corretamente [ʔ] em 41,3% dos casos (n = 499) e [θ] em 7,5% (n = 610). Embora os resultados tenham se mostrado estatisticamente significativos, a autora salienta que as baixas taxas de realização de [θ] podem ter sido enviesadas em virtude do número reduzido de participantes no estudo: havia 16 informantes de proficiência intermediária-baixa e 8 de avançada. De todo modo, ainda que não se possa ter precisão sobre o percentual de acerto dos fones, os resultados indicam que os aprendizes têm altas taxas de erro na produção

das interdentais e que se trata, portanto, de uma dificuldade de pronúncia dos aprendizes brasileiros. Reis (2006) verificou substituições de [?] por [t], [f], [d], [t?], [s] e [t?]; sendo as mais frequentes [t] (45,8%), [t?] (7,5%) e [f] (6,9%); e substituições de [th] por [d], [t?], [d?], [d?], [t?], [?] e [t]; as mais frequentes [d] (85,6%) e [t?] (1,4%)[10].

Trevisol (2010) realizou um estudo sobre a produção da interdental vozeada [th] por professores de inglês. O experimento consistiu da leitura de 20 frases, as quais continham o fone [th] em início e final de palavra. Mesmo nessa população de nível avançado em inglês, as interdentais se mostram como uma dificuldade de pronúncia. Em início de palavra, os sujeitos produziram corretamente [th] em 51,4% das vezes, tendo trocado [th] por [d] nos demais 48,6% casos (n = 220). Em final de palavra, a taxa de acerto verificada foi consideravelmente menor: 26,0% (n = 208). O maior número de substituições deu-se pela correspondente desvozeada da interdental [?], que contabilizou 65,5% das realizações (n = 208); a seguir, o maior número de substituições foi pela oclusiva [d], com 6,0% dos casos (n = 208). Trevisol (2010), também registrou substituições por [v], [f], [t], [t?] e [Ø], no entanto, todas elas são de baixa ocorrência (<1,0%) e podem ser consideradas espúrias. A autora explica a predominância de [?] em final de palavra, em virtude de haver restrições, no PB, para que sons fricativos sejam vozeados em final de palavra. Esse processo será tratado à parte, na Seção 2.1.1.4.4.

No que diz respeito ao [?], tal fone constitui uma aproximante alveolar vozeada e está presente em diversos dialetos do PB. Equivocadamente, a literatura linguística no Brasil convencionou a chamar tal som de “r retroflexo”, embora se trate, em termos articulatórios, de uma aproximante (Rennicke, 2011). A Figura 4 apresenta um mapa da distribuição dos róticos no Brasil, indicando as regiões em que ocorre o [?].

[pic]

Figura 4: Distribuição geográfica dos sons róticos em coda no Brasil - Rennicke (2011) com base em Noll (2008).

Como se nota, a maior parte dos dialetos que contém a aproximante [?] está concetrada na região Centro-Sul do Brasil. Apesar disso, Rennicke (2011) afirma haver estudos que indicam a presença da aproximante [?], em menor concentração, em quase todas as regiões do país. Para os dialetos que possuem a aproximante [?] como parte do inventário fonético, sua percepção e produção na interlíngua não constitui problema, de forma que os aprendizes conseguem, por exemplo, pronunciar car [’k??] e word [’w??d] sem dificuldades.

No entanto, nos dialetos do PB em que [?] não ocorre, os aprendizes têm mais um obstáculo a vencer no aprendizado da língua inglesa. Geralmente, eles acabam por realizar substituições, na interlíngua, mapeando a aproximante [?] em um rótico análogo no PB: [x], [?], [h], [?] ou [?] (Zimmer, 2009).

Osborne (2010) pesquisou a aquisição da aproximante [?] por três brasileiros aprendizes de inglês, com conhecimento de inglês em nível iniciante. O experimento consistiu na leitura de sentenças em voz, as quais continham a aproximante [?] em diversos contextos. Para onset complexos, a exemplo da palavra travel ['træv.?l], [?] foi realizado como [?] em 71,7% dos casos, como [?] em 26,4% e omitido em 1,9% (n = 53). Em posição intervocálica, como America [??mer.?.k?], [?] foi realizado como [?] em 51,7% das vezes e como [?] nos demais 48,3% (n = 29). Em posição de coda, como em park ['p??rk] e war ['w??r], a aproximante [?] apresenta o maior número de variação, sendo realizada ora como [?], [?], [x], [h] ou sendo apagado. Osborne (2010) analisa os dados de coda em duas situações: em meio e final de palavra. No que diz respeito ao meio de palavra, os resultados obtidos com o [?] foram: [h] 57,6%; [?] 18,2%; apagamento 15,1%; [x] 6,1%; e [?] 3,0% (n = 33). Em final de palavra, as realizações são similares, mas não se nota a ocorrência do tepe [?]: apagamento 52,5%; [?] 27,5% [h] 15,0% e [x] 5,0% (n = 40). O autor também avaliou a realização do [h] em inglês, em palavras como huge ['hju?d?], e do tepe [r], como em city [?s??.i]; no entanto, nenhuma variação foi observada, tendo os aprendizes produzido o padrão esperado em 100% dos casos.

8 Falta de aspiração de oclusivas em posição de início de palavra ou sílaba acentuada

Definimos como falta de aspiração de oclusivas a substituição das consoantes [p?], [t?] e [k?] do inglês, em posição de início de palavra ou sílaba acentuada, por suas correspondentes não-aspiradas [p], [t] e [k].

A aspiração é um fenômeno restrito às consoantes obstruintes. Uma consoante é considerada aspirada quando é articulada de modo que, após a fase de explosão dos articuladores, segue-se a liberação de um sopro de ar. Desde Lisker e Abramson (1964), estudos envolvendo comparações entre segmentos vozeados, desvozeados e aspirados têm sido realizados a partir de medidas de voice onset time (VOT). O VOT consiste no intervalo e entre a explosão dos articuladores da consoante e o início do vozeamento da glote. A Figura 5 ilustra as combinações de valores de VOT em uma oclusiva bilabial.

[pic]

Figura 5: Tipos de fonação e valores correspondentes de VOT.

Como se observa, assume-se como ponto de referência, ou ponto zero, a soltura dos articuladores da consoante, a partir disso, calculam-se os valores de VOT. Quando há um intervalo entre a soltura dos articuladores e o início do vozeamento da glote, isto é, VOT > 0, a consoante é classificada como aspirada, no exemplo: [p?]. Quando o vozeamento se inicia imediatamente após a soltura dos articuladores, no caso de VOT = 0, a consoante é desvozeada: [p]. Por fim, quando o vozeamento antecede a explosão, ou seja, quando há pré-sonorização, VOT < 0, a consoante é vozeada: [b].

A relação entre as medidas de VOT e tipo de fonação é dependente de língua, por exemplo, um fone que tenha um determinado valor de VOT pode ser considerado aspirado em uma língua, mas desvozeado em outra. A

., baseada nos dados de Yang (1993), apresenta os valores médios de VOT para as consoantes oclusivas do inglês.

Tabela 1: Valores médios de VOT (em ms) de consoantes oclusivas no inglês (Yang 1993).

[pic]

Yang (1993) opta por reportar os valores de VOT para obstruintes vozeadas na forma A/B, em que A contém a média das amostras com valor negativo de VOT e B as com positivo. O autor argumenta que a diferença de sinal no VOT representa fenômenos diferentes: um VOT < 0 corresponde à pré-sonorização, já um VOT > 0 corresponde à pós- sonorização. Tais fenômenos, de acordo com Yang (1993) devem ser tratados distintamente. A partir Tabela 1, nota-se que, na língua inglesa, o VOT está relacionado ao lugar de articulação da consoante, sendo que o par de alveolares [t?] e [d] apresenta os maiores valores médios, respectivamente 95ms e 20/-91ms. As oclusivas velares [k?] e [g] seguem com 88ms e 32/- 78ms. Por fim, os menores valores médios de VOT são encontrados nas bilabiais [p?] e [b], 77ms e 17/-78ms.

Para o PB, a análise mais extensa de VOT de que temos notícia foi realizada por Klein (1999). Os resultados estão resumidos na Tabela 2.

Tabela 2: Valores médios de VOT (em ms) de consoantes oclusivas no PB (Klein 1999).

[pic]

Como se nota, no PB, as consoantes desvozeadas apresentam menores valores médios de VOT que no inglês, havendo, portanto, menos aspiração. Além disso, a correlação entre o lugar de articulação da consoante e os valores VOT é mais atenuada, a diferença de VOT entre [p] e [t] é estatisticamente insignificante, sendo que apenas a oclusiva velar [k] apresenta algum nível de aspiração. Ao se comparar as medidas de VOT entre os pares do PB e do inglês [p?] vs. [p], [t?] vs. [t], e [k?] vs. [k], é possível ver que as consoantes desvozeadas da língua inglesa apresentam valores bem mais altos de VOT, o que evidencia, de fato, sua aspiração.

Salienta-se que, diferentemente de Yang (1993), Klein (1999) agrupa os valores de VOT positivos e negativos das consoantes vozeadas, de forma que a média apresentada das vozeadas acaba por se tornar menor. Ainda assim, é possível notar que as oclusivas vozeadas do PB apresentam mais pré- sonorização que as do inglês. A maior diferença é verificada na oclusiva velar, no PB, seu valor médio é de -91ms, enquanto no inglês é de 32/-78ms. A seguir a maior diferença é notada na bilabial [b], -87ms vs. 17/-78ms; por fim, os menores

valores são encontrados na alveolar [d], -99 vs. 20/- 91ms. Por conseguinte, conclui-se que as oclusivas vozeadas [b], [d] e [g] do PB possuem mais vozeamento que as do inglês.

Além disso, é possível observar certa sobreposição entre os valores positivos de VOT das oclusivas vozeadas no inglês e das oclusivas desvozeadas no PB. Por exemplo, o valor de VOT positivo de [b] no inglês, 18ms, é bastante similar ao valor de [p] no PB, 17ms. Isso é válido também para os demais lugares de articulação: o valor de VOT da alveolar vozeada [d] do inglês, 20 ms., é muito similar ao da desvozeada [t] do PB, 17 ms.; e o mesmo para as velares [g] e [k], 32ms vs. 38ms. Em outras palavras, as consoantes desvozeadas no PB são articuladas, em certos contextos, de modo muito similar às vozeadas no inglês. Isso pode trazer problemas na inteligibilidade da pronúncia do aprendiz, uma vez que, caso ele transfira esse padrão para a interlíngua, pode, por exemplo, tentar pronunciar [k] e acabar sendo entendido como [g], em palavras como caught ['k??t] e got ['g??t]; coat ['ko?t] e goat ['go?t], etc. Cabe, portanto, ao aprendiz dominar essa diferença, de modo a aumentar a inteligibilidade de sua pronúncia.

Alves (2011) investigou a produção das oclusivas aspiradas [p?], [t?] e [k?] por brasileiros aprendizes de inglês. A autora utilizou valores de VOT na classificação, tendo definido como aspirados os segmentos que apresentavam VOT > 60ms e não-aspirados os que apresentavam VOT < 35ms. A autora observou que [p?] foi realizado sem aspiração pelos aprendizes em 59% das vezes (n = 41), [t?] em 33% das vezes (n = 51) e [k?] em 10% das vezes (n = 96). A diferença no desempenho pode ser explicada em virtude do método de classificação utilizado: apenas duas faixas de valores, VOT < 35ms ou VOT > 60ms; e também em virtude de, no PB, a consoante [k] já possuir maior VOT, dado o lugar de articulação. Entretanto, o estudo contou com um número muito reduzido de participantes (três), de modo que os resultados obtidos devem ser considerados apenas indicativos e não conclusivos.

Prestes (2012) investigou a produção de oclusivas surdas e sonoras por aprendizes brasileiros e falantes nativos de inglês. A autora emprega medidas de VOT na classificação dos segmentos e adota uma postura de tratar o fenômeno em sua gradiência, isto é, considerando-se apenas as medidas de VOT, sem dizer se uma determinada consoante é ou não aspirada. Valores relativos de VOT foram utilizados na análise, mais especificamente, a razão entre a duração do VOT e o tempo total da consoante. Prestes (2012) concluiu que as realizações surdas dos aprendizes apresentaram menor VOT (5,87% VOT/consoante) em relação às dos nativos (2,11% VOT/consoante); e que as vozeadas dos aprendizes apresentam maior VOT (10,15% VOT/consoante) em comparação com as dos nativos (3,89% VOT/consoante) (n = 450). Os resultados indicam, portanto, que brasileiros tendem a realizar [p], [t] e [k], na interlíngua, com menos aspiração que falantes nativos; e [b], [d] e [g] com maior grau

de vozeamento. Ressalta-se, todavia, que o estudou também utilizou um número reduzido de participantes, dois brasileiros aprendizes de inglês e dois nativos.

Schwartzhaupt (2012) analisou o impacto que fatores fonético-fonológicos podem ter sobre os valores de VOT, seu experimento foi conduzido com dez brasileiros aprendizes de inglês e cinco nativos. Foram analisados os seguintes fatores fonético-fonológicos: (i) lugar de articulação da consoante, (ii) qualidade da vogal adjacente e (iii) o número de sílabas da palavra-alvo. Os aprendizes possuíam conhecimento intermediário-avançado ou avançado de inglês. Apesar de o foco do estudo não ser a produção dos aprendizes, mas os efeitos dos contextos fonético-fonológicos no VOT, Schwartzhaupt (2012) observou que os aprendizes foram capazes de realizar, na interlíngua, valores de VOT bem próximos aos dos nativos. Ele salienta o fato de os aprendizes terem conseguido partir de um sistema sem distinção de VOT entre [p] e [t], como é o caso do PB, e terem alcançado, na interlíngua, um sistema em que tal distinção é patente, como é o caso do inglês. Na interlíngua, as realizações de [t?] (77ms) dos aprendizes apresentaram VOT significativamente maior que as de [p?] (61ms), tal como ocorre com falantes nativos.

9 Desvozeamento de obstruintes em posição de final de palavra

O m Xxxxx

10 Vocalização de laterais em final de sílaba

Baratieri (2006) investigou a produção da lateral [l] em coda por um grupo de vinte estudantes de inglês como língua adicional, de proficiência intermediária e avançada. Os resultados obtidos indicaram que a vocalização se trata de um fenômeno gradiente, tendo o autor optado por classificar os dados em três categoriais: (a) segmento parcialmente vocalizado; (b) vocalizado [w]; e (c) não vocalizado [l]. A categoria (a) indica os segmentos para os quais o grau de vocalização não é imediatamente perceptível, de maneira que não há um símbolo IPA correspondente. As produções dos aprendizes apresentaram a seguinte distribuição: 2,7% de [l], 35,5% de [w] e 61,8% de segmentos parcialmente vocalizados (n = 2134). Como se observa, a taxa de produção do padrão esperado [l] foi extremamente baixa, indicando que a vocalização de laterais perdura mesmo em estudantes de proficiência intermediária ou avançada.

11 Quada da nasal em final de sílaba + nasalização da vogal precedente

Kluge e Baptista (2008) estudaram a produção das nasais [m] e [n] em posição final de palavra por um grupo de dez aprendizes de nível intermediário de proficiência. A tarefa consistiu na leitura de sentenças em voz alta. O corpus foi formado por 72 sentenças, 36 contendo [m] e 36 contendo [n], em posição final de monossílabo acentuado do tipo (C)CVC. Os estudantes realizaram a nasal alveolar [n] final esperada em 78,6% dos casos (n = 359), e a nasal bilabial [m] esperada em 63,9% dos casos (n = 357); nos demais, houve apagamento

da nasal e a nasalização da vogal precedente. Testes estatísticos indicaram que a diferença de acerto nas realizações de [m] e [n] é significativa, havendo, portanto, mais dificuldade na aquisição do [m] final pelos aprendizes. As autoras justificam que esse resultado pode se dar em virtude de, no PB, as palavras que terminam em nasal tenderem a ser escritas com (ex.: “fim”, “correm”, “amam”), havendo poucas palavras com (ex.: “hífen”, “pólen”, “abdômen”). Sendo assim, o padrão com é reforçado e o aprendiz apresenta mais resistência para superá-lo na interlíngua.

12 Paragoge da consoante oclusiva velar vozeada [g]

O Xxxxx

13 Assimilação vocálica

O Xxxxx

14 Epêntese interconsonantal (morfema -ed)

O Xxxxx

2.2.2 Why Consider L1 in L2 Teaching?

Lorem ipsum at nusquam appellantur his, ut eos erant homero concludaturque. Albucius appellantur deterruisset id eam, vivendum partiendo dissentiet ei ius. Vis melius facilisis ea, sea id convenire referrentur, takimata adolescens ex duo. Ei harum argumentum per. Eam vidit exerci appetere ad, ut vel zzril intellegam interpretaris.

Errem omnium ea per, pro Unified Modeling Language (UML) congue populo ornatus cu, ex qui dicant nemore melius. No pri diam iriure euismod. Graecis eleifend appellantur quo id. Id corpora inimicus nam, facer nonummy ne pro, kasd repudiandae ei mei. Mea menandri mediocrem dissentiet cu, ex nominati imperdiet nec, sea odio duis vocent ei. Tempor everti appareat cu ius, ridens audiam an qui, aliquid admodum conceptam ne qui. Vis ea melius nostrum, mel alienum euripidis eu.

2.2.3 Common Mispronunciation

Syllable Simplification Lorem ipsum

Consonant Change Lorem ipsum

Deaspiration of Voiceless Plosives in Initial or Stressed Positions Lorem ipsum

Terminal Devoicing in Word-Final Obstruents Lorem ipsum

Delateralization and rounding of lateral liquids in final position Lorem ipsum

Vocalization of final nasals Lorem ipsum

Velar consonantal paragoge Lorem ipsum

Vowel assimilation Lorem ipsum

Interconsonantal epenthesis (-ed and -s morphemes) Lorem ipsum

2.3 Automatic Speech Recognition

2.3.1 The big picture

Basically, all statistical methods of Automatic Speech Recognition (ASR) are dedicated into solving one fundamental equation, which can be described as follows. Let O be a sequence of observable acoustic feature vectors and W be a word sequence, the most likely word sequence $W*$ is given by:

$$W* = \arg\max_{W} P(W|O) \tag{2.1}$$

To solve this equation straightforwadly, one would require a discriminative model, capable of estimating the the probability of W directly from a set of observations O [**?**]. However, HMM are generative models and are not adequate for solving this equation, therefore we apply Bayes' Theorem to Equation 2.1 and end up with:

$$W* = \arg\max_{W} \frac{P(O|W)P(W)}{P(O)} \tag{2.2}$$

As one might notice, we can apply a generative model to calculate the conditional probability term of this equation, that is, the probability of the observation sequence O given a word sequence W, hence $P(O|W)$. At first, it might seem counter-intuitive to conceive a generative model for data analysis, since the data is already available, i.e. O is known before-hand. As **?**] [**?**] points outs, in order to understand how to generative models are used for data analysis, a mental trick is necessary.

[To fix - Fink citation]First one assumes, that the data to be analyzed were generated by a natural process, which obeys similar statistical regularities. Then one tries to reproduce this process with the capabilities of hidden Markov models as closely as possible. If this

attempt is successful, on the basis of the artificial model inferences can be drawn on the real process. On the one hand this may concern the probability for generating the available data. On the other hand the inference on the internal processes within the model is at least probabilistically possible. In particular one can determine the state sequence that generated a certain sequence of outputs with highest probability.

For a single audio input, which we want to decode, the audio is already fixed, so the probability of the observable acoustic feature vectors $P(O)$ is a constant and, therefore, might be discarded. Thus the final fundamental equation is simplified to:

$$W* = \arg\max_{W} P(O|W)P(W) \tag{2.3}$$

$P(O|W)$, the probability of an observable acoustic feature vector given a word sequence, is calculated by an acoustic model. In turn, $P(W)$, the *a priori* probability of words is reckoned by a language model.

2.3.2 Hidden Markov Models and Speech Recognition

Markov models consist a set of mathematical models which are suitable for the statistical description of symbol and state sequences [**?**]. The simplest form of Markov models are Markov chain models, which represent a system with a set of spaces in which transitions from one state to another occur. Within Markov processes, systems are assumed to be memoryless, that is, the conditional probability of future states is only dependent on the present state. To put it another way, Markov models assume that, given a certain system with states and transitions, the current state does not depend upon the sequence of events that preceded it, the so-called Markov property.

HMMs can be formally described as a 5-tuple $\lambda = (Q, O, \Pi, A, B)$, where $Q = \{q_1, q_2, q_3, ..., q_N\}$ is a set of N states. $O = \{o_1, o_2, o_3, ..., o_T\}$ is a set of T observations taken from time $t = 1$ to $t = T$. At each time t it is assumed that the system will be at a specific state q, which is hidden, only the observations are directly visible. $\Pi = \{\pi_i\}$ is a vector with the initial state probabilities, such that

$$\pi_i = Pr(q_i), t = 0 \tag{2.4}$$

$A = [a_{ij}]$ is matrix with the state transition probabilities so that

$$a_{ij} = P(q_t = j|q_{t-1} = i), 1 \leq, i, j \leq N \tag{2.5}$$

and $B = [b_{jt}]$ is a matrix with the emission probability of each state. Assuming a GMM to model the state emission probabilities – the so-called GMM/HMM model; we can define

that, for a state j, the probability $b_j(o_t)$ of generating o_t is given by

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_{js}} c_{jsm} \mathcal{N}(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s} \tag{2.6}$$

where γs is a stream weight, with default value is one, M_{js} is the number of mixture components in state j for stream s, c_{jsm} is the weight of the m^{th} component and $\mathcal{N}(\cdot; \mu_{jsm}, \Sigma_{jsm})$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ, that is

$$\mathcal{N}(o; \mu, \Sigma) = (\sqrt{(2\pi)^n |\Sigma|})^{-e^{-\frac{1}{2}(o-\mu)^T \Sigma^{-1}(o-\mu)}} \tag{2.7}$$

where n is the dimensionality of o.

The following constraints apply:

$$a_{ij} \geq 0 \tag{2.8}$$

that is, the probability of moving from state from any state i to j is not null, and

$$\sum_{j=1}^{N} a_{ij} \geq 1, \forall i \tag{2.9}$$

2.3.3 PB

15 Reconhecimento Automático de Fala (RAF)

O propósito de um reconhecedor de fala é transformar, de forma eficiente e precisa, o sinal acústico da fala em sua contraparte textual (Rabiner & Schafer, 2007). Se cada palavra da língua fosse pronunciada de forma idêntica por todos os falantes e em todos os contextos, a tarefa de Reconhecimento Automático de Fala (doravante RAF) seria algo banal. Mas isso não acontece: a realidade linguística é demasiado variante, quer inter- , quer intrafalantes. Pode-se, por fim, dizer que uma vogal nunca é pronunciada de uma mesma maneira. Furui (2001) sumariza os problemas do RAF em quatro: (i) dificuldades em lidar com coarticulação e redução; (ii) dificuldades de segmentação da fala; (iii) diferenças inter-individuais; e (iv) insuficiência de conhecimento linguístico. A seguir, cada um desses pontos serão discutidos.

A coarticulação é um fenômeno motor que envolve a realização de gestos articulatórios simultâneos ou sobrepostos (Crystal, 2008). Na fala, a coarticulação pode ocorrer de duas formas: antecipatória (left-to-right) ou preservativa (right-to-left). Na coarticulação antecipatória, um som é realizado tomando as características de outro que lhe sucede. Em "pulo", por exemplo, a consoante /p/ é realizada com protrusão labial, [p?], em virtude de a vogal que lhe segue, [u], ser arredondada. Por sua vez, na coarticulação preservativa, um

som é realizado mantendo-se as características articulatórias de outro que lhe antecede. Em dialetos em que não houve a vocalização da lateral, por exemplo, a consoante /l/ da palavra "sal" pode ser realizada de modo velarizado, [?] em vez de [l], por suceder a vogal [a], que é baixa. De fato, a todo o tempo, os sons da fala estão sujeitos à coarticulação. Por tal razão, frequentemente, utilizam-se no RAF unidades fonéticas dependentes de contexto, a exemplo de trifones.

Um trifone é uma representação fonética contextual, que considera, para um dado fone, o fone anterior e o seguinte. A palavra 'fala', por exemplo, pode ser representada, no IPA, pela sequência de fones ['fal?] ou de trifones [#'fa fal al? l?#], especificando-se as articulações secundárias dos fones. No reconhecimento de fala, por razões de codificação em computadores, em geral, os trifones são especificados no formato "L-X+R", em que "X" é um determinado fone, "L" o fone antecedente, e "R" o sucessor. A Figura 6 compara a transcrição da palavra "translate" em fones e trifones.

[pic]

Figura 6: Transcrição da palavra "translate" em fones e trifones de acordo com a convenção do VoxForge (2013).

Quanto às diferenças interindividuais, elas são tamanhas que há até mesmo todo um ramo de investigação da Linguística que se dedica a seu estudo: a Sociolinguística. A Sociolinguística estabelece que a realização linguística é condicionada por diversos fatores sociais: o nível escolar do falante, seu sexo, sua idade, seu estrato social, o lugar onde nasceu e viveu, a situação comunicativa em que está inserido, a formalidade de registro que o contexto demanda, o grau de hierarquia que mantém com seu interlocutor, etc. (Labov, 2008; Weinreich et al., 2006). Para além dos fatores que condicionam a variação, há fenômenos linguísticos que possuem comportamento puramente estocástico: o alçamento vocálico acontece em certas vezes e não em outras, mesmo havendo o mesmo ambiente e considerando- se o mesmo falante.

Não bastassem os fatores sociais, a fala sofre influência também de características anatômicas do indivíduo. A variabilidade linguística começa já quando somos crianças, na fase de desenvolvimento puberal. Das primeiras palavras que balbuciamos até o período de muda vocal, ocorrem grandes mudanças na voz tendo em vista, especialmente, modificações nas configurações das estruturas laríngeas (Guimarães, 2006). Perry et al. (2001) conduziram um estudo sobre vogais produzidas por crianças dos 2 aos 16 anos e verificaram que há diferenças significativas nos valores de frequência fundamental [pic] e nos formantes das vogais [pic]. Ao longo de todo o período de maturação de voz, os valores de frequência dos formantes das vogais [pic] varia significativamente e é distinto para meninos e meninas; após os 12 anos, há grande decréscimo nos valores de frequência fundamental [pic] para os

meninos, de modo que a frequência fundamental [pic] passa também a servir perceptualmente para a distinção do sexo das crianças (Perry et al., 2001) (Figura 7).

[pic]

Figura 7: Comparação dos valores de ??0, F1, F2 e F3 para crianças de ambos os sexos dos 2 aos 16 anos (Perry et al., 2001).

Aos sistemas de RAF, cabe a tarefa de reconhecer, ante toda a variabilidade linguística existente, o que há de invariante na fala. Do ponto de vista computacional, a construção de um reconhecedor de fala pode ser vista como uma tarefa que abrange quatro fases principais: preparação de dados, treino, teste e análise (Young, et al., 2006). A preparação de dados consiste na gravação de arquivos de áudio e texto, em sua anotação, transcrição e pré-processamento, de modo a garantir que a entrada seja compatível com o esperado pelo reconhecedor. O treino constitui a fase em que os dados coligidos são utilizados para criar as componentes do reconhecedor. Em geral, os sistemas de RAF possuem três componentes: (i) um modelo de língua, (ii) um modelo acústico e (iii) um modelo de pronúncia, ou dicionário de pronúncia. Por fim, as fases de teste e análise buscam verificar se os modelos construídos se adéquam à realidade linguística ou ao propósito a que o reconhecedor se dispõe.

16 Representação Digital da Fala

Para desenvolver tecnologias de fala, é preciso antes encontrar formas de codificar, computacionalmente, a informação presente na fala. A fala humana, como constitui um veículo de informação, pode ser vista a partir de uma perspectiva da Teoria Matemática da Comunicação (Shannon, 1948)[11], sendo considerada um sinal acústico. Tecnicamente, sinais são sequencias de estados em um sistema de comunicação que codificam uma mensagem.

Sinais constituem o objeto central de estudo da área de Processamento de Sinal, que busca investigar formas de analisar ou modificar os sinais, no intuito de deles extrair informação ou de adequá-los a um determinado fim (Ingle e Proakis, 2011[12]). Embora a maior parte dos sinais que nos rodeia seja analógica, em muitas vezes, seu processamento é feito não de forma analógica, mas digital. De acordo com Ingle e Proakis (2011)[13], os sistemas de Processamento de Sinal Digital (DSP) são vantajosos, pois:

não necessitam a aquisição de equipamento específico (que são, muitas das vezes, caros), podendo ser desenvolvidos em computadores pessoais;

baseiam-se apenas em operações numéricas de adição e multiplicação, o que lhes dá estabilidade, não sendo necessário calibragem ou padronização, como é comum nos sistemas analógicos;

são altamente adaptáveis, de modo que suas operações podem ser modificadas em tempo real, usualmente, através de técnicas de programação simples.

Um esquema de um sistema de processamento de sinal digital é apresentado na Figura 8.

[pic]

Figura 8: Esquema de um sistema de processamento de sinal digital (Ingle e Proakis, 2011).

Embora a maior parte dos sinais do mundo seja analógica - entre os quais a fala, seu processamento, muitas vezes, dá-se gere sinais analógicos,

Embora a maior parte dos sinais com que

O processamento de sinais pode ser feito de forma analógica ou digital.

A área do conhecimento responsável por tal busca é o Processamento de Sinal.

A área responsável por esse

Do ponto de vista da Teoria da Informação, sinais são sequencias de estados em um sistema de comunicação que codificam uma mensagem.

(Ingle e Proakis, 2010)

17 Tipos de Reconhecedores de Fala

Reconhecedores Automáticos de Fala podem ser agrupados em três categoriais, que se dividem quanto à tarefa de reconhecimento que desempenham: i) reconhecimento de palavras isoladas; ii) de sentenças pré-estabelecidas; iii) reconhecimento de fala contínuo de grande vocabulário (RFCGV)[14] (Rabiner, 1997).

Os sistemas de reconhecimento de palavras isoladas são utilizados, por exemplo, por centrais telefônicas, nas Unidades de Resposta Audível (URA), em menus do tipo: "Fale 'um' para ser redirecionado ao setor de cancelamento" ou "Para conversar com um de nossos atendentes, fale 'atendente' ". Reconhecedores de fala do segundo tipo são mais robustos que os do segundo, sendo capazes de reconhecer sentenças pré-definidas ou provenientes de uma gramática pré-estabelecida. Um exemplo de aplicação deste tipo são os módulos de comando por voz de computadores, celulares e os sistemas hands-free em carros, em que se pode dizer sentenças do tipo "ligar o rádio" ou "Siri, google 'speech recnognition' "e o comando é reconhecido. Já os sistemas de reconhecimento de fala contínuo de grande vocabulário são os reconhecedores mais abrangentes, sendo capazes de processar a fala espontânea do usuário. Sistemas de diálogo por voz, como os How May I Help You (HMIHY), e sistemas de ditado em editores de textos são desse último tipo.

A seguir, será discutida apenas a arquitetura dos sistemas de RFCGV, primeiro, porque consistem nos sistemas de reconhecimento com a arquitetura mais complexa e, segundo, porque é o tipo de reconhecedor que se pretende desenvolver neste projeto.

18 Arquitetura Básica de um Reconhecedor de Fala Contínuo de Grande Vocabulário (RFCGV)

O paradigma majoritário em sistemas de RAF é estocástico, destacando-se, especialmente, a utilização de Modelos Ocultos de Markov, ou Hidden Markov Models (HMM) (Huang, et

al., 2001). Em tais modelos, a tarefa de reconhecimento é considerada a partir da metáfora do canal ruidoso, ou noisy-channel (Jurafsky & Martin, 2009). O sinal acústico, que constitui a entrada no sistema, é visto como uma deformação da mensagem original, isto é, da sequência de palavras pretendida pelo falante, após passar por um canal com ruído. Assim, o reconhecimento se torna uma tarefa de decodificação, isto é, trata de como recuperar a mensagem original a partir do sinal acústico "ruidoso". Matematicamente, isso corresponde a estimar, considerando-se uma língua [pic], para uma sequência de palavras [pic], qual é a sequência [pic] mais provável, dado conjunto de estados acústicos observáveis [pic]:

[pic]

Todavia, não é possível calcular [pic] diretamente, sendo necessário aplicar-se o Teorema de Bayes, de modo a obter-se:

[pic]

Como a propósito é buscar a sequência de palavras mais provável para um conjunto já dado de estados acústicos, [pic] se repete a cada cálculo, de maneira que pode ser considerado uma constante de normalização, e a equação pode ser simplificada para:

[pic]

Essa equação fundamenta a base dos sistemas de RAF estocásticos e possui estreita relação com a arquitetura que é por eles compartilhada. Basicamente, os sistemas de RAF contínuo com grande vocabulário possuem três módulos: (i) um modelo de língua, (ii) um modelo acústico e (iii) um modelo, ou dicionário de pronúncia. O modelo de língua é utilizado para estimar [pic], a probabilidade a priori da sequência de palavras. Já o modelo acústico é utilizado para calcular [pic], a verossimilhança da observação. Por fim, o dicionário de pronúncia serve como uma ponte entre o modelo de língua e o modelo acústico, uma vez que possui as palavras que compõem o léxico do reconhecedor, transcritas em forma ortográfica e fonética. A Figura 9 ilustra a arquitetura básica de um sistema de RAF.

[pic]

Figura 9: Arquitetura básica de um Reconhecedor Automático de Fala.

O modelo acústico processa o sinal acústico da fala, de modo a inferir quais são os segmentos sonoros que a compõem, usualmente, empregando fones ou trifones. Em reconhecedores de base em HMM, essa tarefa é feita estimando-se os estados acústicos observados mais prováveis, bem como suas probabilidades de transição. Já o modelo de pronúncia provê a correspondência entre sequências de fones e as palavras da língua. No exemplo, tal modelo mapeia a sequência de fones [fal?] na palavra "falo". O modelo de língua, por sua vez, estima as ordenações de palavras mais prováveis na língua.

2.3.4 A Brief History

Although the task of recognizing words from speech might seem apparently simple beforehand, (after all humans start doing it with as little as four months! (XXXX INSERT CITATION)), the issue is actually very complex one. Over the years, many methods have been proposed to attempt to solve the problem of speech recognition. However until now no solution has been found and machines are still a very long way from performing like humans. Table 2.25 presents a comparison between the performance of humans and machines in some recognition tasks.

Table 2.25 Word error rate comparisons between human and machines on similar tasks [21].

Tasks	**Voc. size**	**Humans**	**Machines**
Connected digits	10	0.009%	0.720%
Alphabet letters	26	1%	5%
Spontaneous telephone speech	2,000	3.8%	36.7%
WSJ with clean speech	5,000	0.9%	4.5%
WSJ with noisy speech (10-db SNR)	5,000	1.1%	8.6%
Clean speech based on trigram sentences	20,000	7.6%	4.4%

As one may observe, humans outperform machines in almost every task, specially the more complex ones. Humans' are indeed the topline for the speech recognition task, the uttermost dream of each speech scientist alive is to build a system capable of performing similarly to humans. Although this dream is somewhat near for rather simple tasks like connected digits, for other contexts a long path yet lies ahead.

Recognizing spontaneous speech is still a huge barrier for machines, as can be seen from the huge difference in the spontaneous telephone corpus: whereas humans had a WER of 3.8%, for machines this rate is up to 36.7%. Such result is mainly due to linguistic variability. Language varies not only among speakers (the so-called inter-speaker variability), but also within the same speaker (intra-speaker) [1].

Considering inter-speaker differences, factors such as gender, age, social, and regional origin, health and emotional state might have a huge impact on the speech signal [1]. Sociolinguistics has long known that gender affects language usage, in fact, men and women tend to use different language constructions. In her seminal paper in the field, Lakoff [23] [23] found that, in women's speech, strong expressions of feeling are avoided, uncertainty is

favored, and means of expression in regard to subject-matter deemed “trivial” to the “real” world are elaborated.

Put aside social aspects, men, women and children’s speech are also contrasting because of morphological differences in their vocal tract. Sex and development influence body size, and there is a strong correlation between vocal tract length and body size (either height or weight); in addition to this, the relative proportions of men and women’s oral and pharyngeal cavity are unlike [16]. Figure 2.5 presents a comparison between height and vocal tract length for men, women and children. Figure 2.6 presents a model of the vocal tract morphology considering age.

Fig. 2.5 Height (cm) versus vocal tract length (mm) [16].

Fig. 2.6 Averaged vocal tract morphology [16].

As one can observe, men’s vocal tract are longer than women’s, followed and obviously children. These morphology differences affect the speech signal thorougly, specially in what concerns to the Fundamental Frequency (F0). F0 can be defined as the lowest frequency in the signal counting from zero. Figure 2.7 compares the F0 values between male and females considering aging.

Fig. 2.7 F0 and pitch sigma versus age for males and females [3].

One can notice from Figure 2.7, that no difference is found between male and female voice at a very young age. In fact, boys and girls have roughly the same F0 values. However, when they reach puberty, differences begin to appear. This period is commonly called the voice mutation or voice change, when the F0 for male voice has huge drop, while for female voice the drop is quite small. In terms of perception, this is period when the male voice lowers and geets deeper.

XXXXXXXXXXXXXXXXXXX EXTEND INTER-SPEAKER VARIABILITY XXXXXXXXXXXXX DESCRIBE INTRA-SPEAKER VARIABILITY

But let’s get back to Table 2.25. It is interesting to notice that humans performed better in all speech recognition tasks, but one: “Clean speech based on trigram sentences”. This one task consists of recognizing sentences which were randomly generated using the WSJ trigram language model. Therefore, humans had no advantage over machines in what concerns to syntactic or semantic knowledge. This result highlights one of the most important feature

of human hearing, that is, we make a large use of syntactic, semantic and also pragmatic information in order to understand speech. While hearing do not take into account simply the acoustic signal, but the whole context.

Language is a social tool, which aims at successfull interaction. When someone steps into a snack bar and orders an [aɪsˈkrim], the vendor has no doubt that this sequence of phones refers to "ice cream" and not "I scream", albeit they are pronounced exactly the same. Although such sequence of phones might be ambiguous in the phonetic level, it is not in higher linguistic levels, such as syntax (the verb "order" is usually followed by a noun), semantics (the object of order has to be something purchasable) and pragmatics (one does not buy his own shout!).

2.3.5 HMM-based Speech Recognition

HMM is the most widespread and successfull paradigm in ASR. When HMM were first applied to speech recognition in the late 70's, they were completely revoluationary. Up until recently Deep Neural Network (DNN) seem to be next prominent paradigm in ASR. HMM have been applied to ASR since the late 70's, and they have gathered the best results until recently.

A HMM is a statistical Markov model in which the states are assumed to be hidden, i.e. they are not directly visible, only the state's outputs are observable. Each state has a probability distribution over the possible output tokens, in such a way that output generated by the HMM states provides some information about the hidden sequence of states which was traversed.

2.3.6 Feature Extraction

Preambulus Feature extraction is an importart part of speech recognition systems. The feacture extraction phase is responsible for identifying or enhancing the components of the signal that are relevant for recognizing speech sounds, while discarding or diminishing the effect of unuseful information, such as background noise. With respect to speech parameterization, Mel Frequency Cepstral Coefficients (MFCC) are definitely the standard. MFCC have been widely used in ASR systems for almost three decades [10], they are present on the many important speech recognition toolkits, such as Hidden Markov Model Toolkit (HTK), Sphinx, RASR (RASR) and Kaldi. Before we go into further details about these features it is interesting to give a little background about speech recording and coding.

Speech is recorded by using a microphone – nothing new so far! Despite the many types of available microphones (condenser, capacitor, pyezoeletric, laser, etc.) its design remains

basically the same as the carbon microphone invented by David Hughes two centuries ago [30]. A microphone is simply an acoustic-to-electric sensor, which converts variations in air pressure (that is, sound) into an electrical signal. Microphones have a very thin membrane, called diaphragm, which vibrates when struck by sound waves. When the diaphragm vibrates, it puts to move a sensitive capsule attached to it, that converts its movement into electrical pulses. Most of the current microphones are dynamic, which means that their capsule consist of .... (XXX VER WIIPEDIA)

After capturing speech through a microphone, one usually wants to store it for later access. In order to store speech digitally on a computer, a coding scheme is mandatory. In the literature, many coding schemes have been proposed, such as linear PCM, μ-law, A-law PCM, APCM, DPCM, DM, and ADPCM [21]. The details of each type of speech coder is beyond the scope of this dissertation, the reader can find an description of each scheme in Huang et al. [21] [21] or Furui [17] [17]. Following we will give a brief discussion of linear PCM, which is the standard way of storing audios in digital format.

Pulse Code Modulation (PCM) is a type of analog-to-digital conversion, which constitutes the basis of the WAV digital audio format, together with other lossless formats such as AIF and AU. [3]. PCM coding is based on two properties: (i) a sampling rate of the audio and a (ii) bit depth. The sampling rate determines the number of audio samples that are taken per second from the signal, in turn the bit depth is the number of bits of information in each audio sample. Both values must be constant and should be defined prior to recording (actually coding) an audio. The sampling and the bit depth are closely related to the audio quality, that is, the higher the sampling and the depth the better the fidelity of the digital audio to the analog speech signal. Picture XXX presents an example of a linear PCM representation, at different sampling rates and bit depths, of an audio containing the utterance “Speech recognition”.

Linear PCM assumes that the discrete signal $x[n]$ is bounded, that is,

$$|x[n]| \leq X_{max} \tag{2.10}$$

and that the quantization step Δ is uniform for all consecutive levels of x_i

$$x_i - x_{i-1} = \Delta \tag{2.11}$$

Assuming a binary code, the number of levels which can be represented by PCM is $N = 2^B$, where B is the bit depth, this constitutes the audio resolution. According to [21],

[3] Other types of popular audio files which use lossy data compression, such as MP3, WMA, OGG or AAC (a format common to DivX videos) do not use PCM. Instead

speech could be represented in an intelligible way by using 7 bits, however, in practice, applications use values no lower than 11 bits to guarantee communication efficiency. For instance, CDs makes use of 16-bit linear PCM, whereas DVD-Audio and Blu-Ray discs can support up to 24-bit.

Although linear PCM files are able to carry all the necessary auditory information – after all we are able to listen to them and recognize the speech, the music or the noise recorded in them; they are not useful for speech recognition purposes. This occurs because, from the phonological point of view, very little can be said based on the waveform itself [32]. Consider, for instance, the two combinations of 100 Hz, 200 Hz and 300 Hz sine waves, shown in Figure 2.8, which differ only with respect to the relative timing.

Fig. 2.8 Two complex waveforms generated by the same three pure tone 100 Hz, 200 Hz and 300 Hz sine waves, differing only with respect to their relative timing [22].

As one might notice, disregard of being composed by the same pure tones, the complex waves shown in Figure 2.8 are completely distinct from one another. This happens because the waveform is influenced by phase shifts (also known as phase offsets). Therefore in-phase and out-of-phase waves (Figure 2.9 and Figure 2.10) are represented differently, and this adds too much variability to the waveform, in such way that the signal waveform becomes unsuitable for human analysis and consequently for being used as a raw input in ASR systems.

Fig. 2.9 Example of in-phase waves.

Fig. 2.10 Example of out-of-phase waves.

Another way of representing the audio information, which is more meaningful for human reading or computer analysis is through short-term spectrum. Short-term spectra are obtained by applying a Discrete Time Fourier transform to a windowed signal. At first, the signal is divided into uniformly-spaced periods with a sliding window. For speech recognition, usually the window size is defined as 25 ms, with a frame shift of 10 ms, audio information is extracted every 10 ms with 15 ms of overlapping among adjacent frames [21]. Figure 2.11 contains an example of a windowing process (in this case, with 50% overlapping).

These windows values are based on two assuptions: (i) that within 25 ms the signal is stationary, i.e. the phonatory system is not moving; (ii) that at least a period of each relevant speech frequency will be captured by this window this windows, that is no relevant are

Fig. 2.11 Illustration of an original audio recording (the upper waveform) divided into two offset sequences of analysis windows (two lower waveforms) with 50% overlapping frames [25]

After windowing the signal a Fourier transform is applied into each window so as to obtain a series of frequency spectra, i.e. a series of representation of the signal in the frequency domain instead of the time domain. As can be noticed in Figure 2.11, since the frame shift is smaller than the window size, the windowing process extracts many redundant information. The intention for doing this will be made afterwards, when we give further details of the Fourier transform. Such transform is based on the Fourier theorem, which states that any periodic waveform can be approximated as closely as desired as the sum of a series of pure sine waves. In other words, the Fourier transform is able to analyse a short-term of the signal, containing a complex wave, and to output which are and what is the amplitude of the pure tones which form this complex wave.

Feature extraction must then be performed in stored audio files in order to extract relevant information from the waveform and discard redundant or unwanted signal characteristics. As already mentioned before, the two most traditional techniques for speech feature extraction, over the past decades, have been the MFCC [10] and the Perceptual Linear Prediction (PLP) [20]. Both parameterization methods are based on the short-term spectrum of speech. For speech recognition purposes, MFCC features usually show better performance when compared to PLP, for this reason in this thesis we are only going to present MFCC features [**? ?**].

2.3.7 MFCC Features

MFCC is a type of speech parameterization is the result of a cosine transform of the logarithm of the short-term energy spectrum expressed over a mel scale [10]. MFCC features tries to reduce the feature dimensionality of a sound Fourier spectrum, by applying some concepts of Psychoacoustics and Psychophysics in order to the extract a vector with relevant values from the spectrum. The aim is to represent speech data in a compressed format, by eliminating information which are not pertinent to the phonetic analysis and to enhance the aspects of the signal which contribute to the detection of phonetic differences [10].

From Psychoacoustics, MFCCs use the notion that humans do not perceive frequency through a linear scale, but through a scale which resembles to be linear-spaced in frequencies

below 1000 Hz and logarithmic in frequecies above 1000 Hz[4], the so-called mel scale (named after *mel*ody). The scale is based on experiments with simple tones in which individuals are required to separat frequency values into four equal intervals or to adjust the frequency of a stimulus to be half as high as another reference tone [21]. The reference point between mel scale and a linear frequency scale is 1000 mels, which correspond to a 1000 Hz tone, 40 dB above the absolute threshold of hearing. Since it was first introduced by Stevens et al. [36] [36], the scale has been revisited many times [37], but a common formulation, according to Huang et al. [21] [21] is:

$$M(f) = 1125 * ln(1 + f/700) \tag{2.12}$$

where f is the input frequency in Hz. The scale is plotted Figure 2.12.

Fig. 2.12 Mel scale versus a linear frequency scale.

and also of the physics of speech, such as the fact that human like the these systems often have well defined overtones that are harmonic – which is why the MFCCs use the FFT of the FFT)

2.3.8 Dealing with Noisy Data

One of the central problems in ASR is how to deal with noisy audio data. It is long known that the performance of speech recognition systems greatly degrate when the environmental or the recording conditions are not controlled, thus allowing unwanted residual sounds to appear in the signal. In acoustics, any type of sound that is not the one you are willing to analyze is considered noise. As a result from this, in speech recognition, the hiss of a fan, the buzz that a computer cooler makes, car horns on the street and so on are all regarded as noise. Even someone's voice can be regarded as noise. Consider, for instance, that you are trying to recognize John's speech in an application, however Mary is close to him talking on the phone, to the extent that traces of her voice are added to the signal. In this scenario, Mary's voice is actually noisy data, since it is undesirable for the given purpose.

[4]This is not entirely true. As shown by Umesh et al. [37] [37], in fact, there are no two distinguishable regions in terms of statistical significance. But the idea that we perceive low frequencies better than high ones still hold.

2.3.9 Types of Speech Recognition Systems

Errem omnium ea per, pro UML congue populo ornatus cu, ex qui dicant nemore melius. No pri diam iriure euismod. Graecis eleifend appellantur quo id. Id corpora inimicus nam, facer nonummy ne pro, kasd repudiandae ei mei. Mea menandri mediocrem dissentiet cu, ex nominati imperdiet nec, sea odio duis vocent ei. Tempor everti appareat cu ius, ridens audiam an qui, aliquid admodum conceptam ne qui. Vis ea melius nostrum, mel alienum euripidis eu.

2.3.10 The Architecture of a Large Vocabulary Continuous Speech Recognition System

Non vices medical da. Se qui peano distinguer demonstrate, personas internet in nos. Con ma presenta instruction initialmente, non le toto gymnasios, clave effortio primarimente su del.

Chapter 3

Copy of the articles

Using a hybrid approach to build a pronunciation dictionary for Brazilian Portuguese

Gustavo Mendonça, Sandra Aluisio

Instituto de Ciências Matemáticas e de Computação
University of São Paulo, Brazil
gustavom@icmc.usp.br, sandra@icmc.usp.br

Abstract

This paper describes the method employed to build a machine-readable pronunciation dictionary for Brazilian Portuguese. The dictionary makes use of a hybrid approach for converting graphemes into phonemes, based on both manual transcription rules and machine learning algorithms. It makes use of a word list compiled from the Portuguese Wikipedia dump. Wikipedia articles were transformed into plain text, tokenized and word types were extracted. A language identification tool was developed to detect loanwords among data. Words' syllable boundaries and stress were identified. The transcription task was carried out in a two-step process: i) words are submitted to a set of transcription rules, in which predictable graphemes (mostly consonants) are transcribed; ii) a machine learning classifier is used to predict the transcription of the remaining graphemes (mostly vowels). The method was evaluated through 5-fold cross-validation; results show a F1-score of 0.98. The dictionary and all the resources used to build it were made publicly available.

Index Terms: pronunciation dictionary, grapheme to phoneme conversion, text to speech

1. Introduction

In many day-to-day situations, people can now interact with machines and computers through the most natural human way of communication: speech. Speech Technologies are present in GPS navigation devices, dictation systems in text editors, voice-guided browsers for the vision-impaired, mobile phones and many other applications [1]. However, for many languages, there is a dire shortage of resources for building speech technology systems. Brazilian Portuguese can be considered one of these languages. Despite being 6th most spoken language in the world [2], with about 200 million speakers, speech recognition and speech synthesis for Brazilian Portuguese are far from the current state of the art [3]. In this paper, we describe the method employed in building a publicly available pronunciation dictionary for Brazilian Portuguese which tries to diminish this scarcity.

The dictionary makes use of a hybrid approach for grapheme to phoneme conversion, based on both manual transcription rules and machine learning algorithms, and aims at promoting the development of novel speech technologies for Brazilian Portuguese. Hybrid approaches in grapheme to phoneme conversion have been applied successfully to other languages [4][5][6][7]. They have the benefit of taking advantage from both knowledge-based and data-driven methods. We propose a method in which the phonetic transcription of a given word is obtained through a two-step procedure. Its primary word list derives from the Portuguese Wikipedia dump of 23rd January 2014. We decided to use Wikipedia as the primary word list for the dictionary for many reasons: i) given its encyclopedia nature, it covers wide-ranging topics, providing words from both general knowledge and specialized jargon; ii) it contains around 168,8 million word tokens, being robust enough for the task; iii) it makes uses of crowdsourcing, lessening author's bias; iv) its articles are distributed through Creative Commons License. Wikipedia articles were transformed into plain text, tokenized and word types were extracted.

We developed a language identifier in order to detect loanwords among data. It is a known fact that when languages interact, linguistic exchanges inevitably occur. One particular type of linguistic exchange is of great concern while building a pronunciation dictionary, namely, non-assimilated loanwords [8]. Non-assimilated loanwords stand for lexical borrowings in which the borrowed word is incorporated from one language into another straightforwardly, without any translation or orthographic adaptation. These words represent a problem to grapheme-to-phoneme (G2P) conversion since they show orthographic patterns which are not predicted in advance by rules or which are too deviant to be captured by machine learning algorithms. Many algorithms have been proposed to address Language Identification (LID) from text [9][10][11][12]. Since our goal is to detect the language of single words, we employed n-gram character models in the identifier, given its previous success in dealing with short sequences of characters.

Brazilian Portuguese Phonology can be regarded as syllable and stress-driven [13]. In fact, many phonological processes in Brazilian Portuguese are related to or conditioned by syllable structure and stress position [14]. Vowel harmony occurs in pretonic context [15], posttonic syllables show a limited vowel inventory [13], nasalization occurs when stress syllables are followed by nasal consonants [16], epenthesis' processes are triggered by the occurrence of non-allowed consonants in coda position [17] and so on and so forth. Therefore, detecting syllable boundaries and stress is of crucial importance for G2P systems, in order to achieve correct transcriptions. Several algorithms have been proposed to deal with the syllabification in Brazilian Portuguese. However most of them were not extensively evaluated nor were made publicly available [18] [19] [3] [20]. For this reason, we implemented our own syllabification algorithm, based directly on the rules of the last Portuguese Language Orthographic Agreement [21].

Word types recognized as belonging to Brazilian Portuguese by the language identifier were transcribed in a two-step process: i) words are submitted to a set of transcription rules, in which predictable graphemes (mostly consonants) are transcribed; ii) a machine learning classifier is used to predict

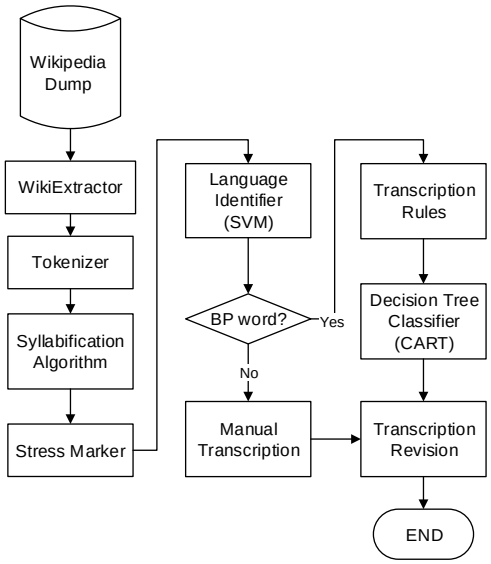


Figure 1: *System architecture for building the pronunciation dictionary.*

the transcription of the remaining graphemes (mostly vowels). All the data were subsequently revised. Figure 1 summarizes the method.

2. Method

2.1. Primary Word List

We used the Portuguese Wikipedia's dump of 23rd January 2014 as the primary word list for the pronunciation dictionary. In order to obtain plain text from the articles, we employed WikiExtractor [22]; it strips all the MediaWiki markups and metadata forms. Afterwards, texts were tokenized and unique words types extracted. The Portuguese Wikipedia has about 168,8 million word tokens and 9,7 million types, distributed among 820,000 articles. With the purpose of avoiding misspellings, URLs and other spurious data, only words with frequency higher than 10, which showed neither digits nor punctuation marks were selected.

2.2. Language Identifier

A Language Identifier module was developed in order to detect loanwords in the pronunciation dictionary. The Identifier consists of a Linear Support Vector Machine Classifier [23] and was implemented in Python, through Scikit-learn [24]. It was trained on a corpus made of the 200,000, containing 100,000 Brazilian Portuguese words and 20,000 words of each of the following languages: English, French, German, Italian and Spanish. All of these words were collected through web crawling News' sites and were not revised. We selected these languages because they are the major donors of loanwords to Brazilian Portuguese [25]. From these words we extracted features such as initial and final bi- and trigraphs; number of accented graphs, vowel-consonant ratio; average mono-, bi- and trigraphs probability; and used them to estimate the classifier. Further details can be found in the website of the Project[1]. After training, we applied the classifier to the Wikipedia word list with the purpose of identifying loanwords among data. The identified loanwords were then separated from the rest of words for later revision, i.e. they were not submitted to automatic transcription.

2.3. Syllabification algorithm and stress marker

Our syllabification algorithm follows a rule-approach and is based straightforwardly on the syllabification rules described in the Portuguese Language Orthographic Agreement [21]. Given space limitations, rules were omitted from this paper as they can be found in the website of the project, along with all the resources developed for the dictionary. As for the stress marker, once the syllable structure is known in Brazilian Portuguese, one can predict where stress falls. Stress falls:

1. on the antepenultimate syllable if it has an accented vowel <á,â,é,ê,í,ó,ô,ú>;
2. on the ultimate syllable if it contains the accented vowels <á,é,ó> or <i,u>; or if it ends with one of the following consonants <r,x,n,l,z>;
3. on the penultimate syllable otherwise.

2.4. Transcriber

The transcriber is based on a hybrid approach, making use of manual transcription rules and an automatic classifier, which builds Decision Trees. Initially, transcription rules are applied to the words. The rules covers not all possible graphemes to phoneme relations, but only those which are predictable by context. The output of the rules is what we called the intermediary transcription form. After obtaining it, a machine learning classifier is applied in order to predict the transcription of the remaining graphemes. Figure 2 gives an example of the transcription process.

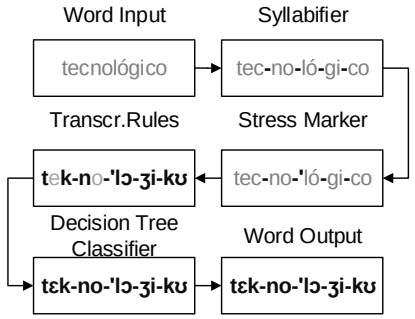


Figure 2: *Example of the transcription procedure – in grey: graphemes yet to be transcribed; in black: graphemes already transcribed.*

The rules' phase has two main goals: guarantee the correct transcription of certain predictable graphemes (mostly consonants) and also ensure the alignment between graphemes and phones for the classifier. They were set in order to avoid overlapping and order conflicts. Long sequences of graphemes, such

[1]http://nilc.icmc.usp.br/listener/aeiouado

as triphthongs, contextual diphthongs and general diphthongs are transcribed first (e.g. <x-ce>→[-se]). Then graphemes involving phones that undergo phonological processes are transcribed (e.g. <ti>→[tʃi], <di>→[dʒi]). After that, several contextual and general monophones are transcribed (e.g. <#x>→[ʃ], <#e-x>→[#e-z]).

On what regards to the classifier, it was developed primarily to deal with the transcription of vowels. In Brazilian Portuguese, vowels have a very irregular behavior, specially the mid ones. Therefore the relations between the vowels' graphemes and their corresponding phonemes are hard to predict beforehand through rules. Consider, for instance, the words "teto" *(roof)* and "gueto" *(ghetto)*; both are nouns and share basically the same orthographic environment. However the former is pronounced with an open "e" ['tɛ.tʊ] and the latter with a closed one ['ge.tʊ]. The classifier employs Decision Trees, through an optimised version of the CART (Classification and Regression Trees) algorithm and was implemented in Python, by means of the Scikit-learn library [24].

The algorithm was trained over a corpus of 3,500 words phonetically transcribed and manually revised, with a total of 39,934 instances of phones. The feature extraction happened in the following way. After reviewing the data, we obtained the intermediary transcription form for each of these words and aligned them with the manual transcription. Then, we split the intermediary transcription form into its corresponding phones and, for each phone, we extracted the following information: i) the phone itself; ii) 8 previous phones; iii) 8 following phones; iv) the distance between the phone and the tonic syllable; v) word class – parts of speech; v) the manually transcribed phone. We considered a window of 8 phones in order deal with vowel harmony phenomena. By establishing a window with such length, one can assure that pretonic phones will be able to reach the transcription of the vowels in the stressed syllable. The classifier was applied to all 108,389 words categorized as BP words by the Language Identifier module, all of them were cross-checked by two linguists with experience in Phonetics and Phonology.

3. Results

The Portuguese Wikipedia has about 168,8 million word tokens and 9,7 million types, distributed among 820k articles. After applying the filters to the data, i.e. words with frequency higher than 10, with no digits nor punctuation marks, we ended up with circa 238k word types, representing 151,9 million tokens. Table 1 describes the data.

Table 1: *Portuguese Wikipedia Summary – Dumped on 23rd January 2014.*

	Word Tokens	**Word Types**
Wikipedia	168,823,100	9,688,039
Selected	151,911,350	238,012
% Used	90.0	2.4

The selected words covers 90,0% of the Wikipedia content. Although the number of selected word types seems too small at first glance, one of the reasons is that 7,901,277 of the discarded words were numbers (81,5%). The remaining discarded words contained misspellings (*dirijem-se* – it should be *dirigem-se*), used a non-Roman alphabet (*λόγω*), were proper names (*Stolichno*, *Zé-pereira*), scientific names (*Aegyptophitecus*), abbreviations or acronyms (LCD, HDMI).

As for the language identifier, we trained and evaluated it with the 200,000 words multilingual corpus. The corpus consists of 100,000 Brazilian Portuguese words and 20,000 words from each of the following languages: English, French, German, Italian and Spanish. All of these words were collected through web crawling News' sites and were not revised. The results obtained for the identifier, through 5-fold cross validation are described in Table 2.

Table 2: *Results from the Language Identifier module – Training Phase.*

	Precision	**Recall**	**F1-score**	**Support**
BP words	0.85	0.89	0.87	100,000
Foreign Words	0.88	0.84	0.86	100,000
Avg/Total	0.86	0.86	0.86	200,000

The classifier showed an average F1-score of 0.86. Although such result is not as good as we expected – some authors reported 99% by using similar methods with trigrams probability, the relatively low F1-score can be explained given the nature of the data. In most language identifiers, the input consists of texts or several sentences, in other words, there is much more data available for the classifier. Since we are working with single words, the confusion of the model is higher and the results are, consequently, worse. Additionally, because the word list used to train the identifier was not revised, there is noise among the data. After training and evaluating the classifier, we applied it to the selected word list derived from the Wikipedia, in order to detect loanwords. Table 3 describes the results gathered.

Table 3: *Results from the Language Identifier module – Wikipedia word list.*

	Wikipedia word list
BP words	108,370 (46%)
Foreign Words	129,642 (54%)
Total	238,012

As one can observe, although we established a frequency filter to avoid spurious words, many loanwords still remain. More than half of the word list selected from Wikipedia consists of foreign words. Notwithstanding that, the list of Brazilian Portuguese words is still of considerable size. For instance, the CMUdict [26], a reference pronunciation dictionary for the English language, has about 125,000 word types.

Concerning the syllabification algorithm and the stress marker, we did not evaluate them in isolation, but together with the transcriber since the rules for each of these modules are intertwined. That is to say the transcription rules are strictly dependent on the stress marker module and the syllable identifier. Besides, the Decision Tree Classifier is built upon the output of the transcription rules, so it is entirely dependent on it. The Decision Tree Classifier was trained over a corpus of 3,500 cross-checked transcribed words, containing 39,934 instances of phones. We analyzed its performance through 5-fold cross validation, the results for each individual phone are summarized in Table 4.

As it can be seen, the method achieved very good results, with a F1-score of 0.98. Many segments were transcribed with 100% accuracy, most of them were consonants. As it was expected, the worst results are related to mid vowels [ɛ, e, ɔ, o],

Table 4: *Results from the Transcriber – Training Phase.*

	Precision	Recall	F1-score	Support
syl. boundary	1.00	1.00	1.00	9099
stress	1.00	1.00	1.00	3507
p	1.00	1.00	1.00	760
b	1.00	1.00	1.00	357
t	0.99	0.99	0.99	1135
d	0.99	0.99	0.99	1148
k	0.99	0.99	0.99	978
g	1.00	1.00	1.00	298
tʃ	0.98	0.98	0.97	450
dʒ	0.96	0.96	0.96	243
m	1.00	1.00	1.00	668
n	1.00	1.00	1.00	556
ɲ	1.00	1.00	1.00	69
f	1.00	1.00	1.00	311
v	1.00	1.00	1.00	531
s	0.98	0.98	0.98	2309
z	0.93	0.94	0.93	416
ʃ	0.84	0.84	0.84	138
k.s	0.72	0.64	0.66	41
ʒ	1.00	1.00	1.00	196
l	1.00	1.00	1.00	682
ʎ	1.00	1.00	1.00	58
ɾ	1.00	1.00	1.00	1388
h	0.98	0.99	0.99	737
ɦ	0.97	0.92	0.94	169
w	0.97	0.98	0.97	441
w̃	0.98	0.99	0.99	309
j	0.97	0.95	0.96	223
j̃	0.95	1.00	0.98	110
a	1.00	1.00	0.99	2316
ə	0.99	0.99	0.99	1093
ɛ	0.65	0.68	0.66	275
e	0.93	0.91	0.92	1779
i	0.98	0.99	0.98	2073
ɪ	0.97	0.97	0.97	365
ɔ	0.69	0.75	0.71	220
o	0.93	0.92	0.93	1112
u	0.96	0.96	0.96	488
ʊ	1.00	1.00	1.00	1033
ã	1.00	1.00	1.00	719
ẽ	0.96	0.97	0.97	497
ĩ	0.99	0.99	0.99	274
õ	0.97	0.96	0.97	299
ũ	0.94	0.92	0.93	64
Avg/Total	0.98	0.98	0.98	39934

specially mid-low vowels, [ɛ] showed a F1-score 0.66 and [ɔ] of 0.71. It can be the case that since the grapheme context is the same for [ɛ, e] and [ɔ, o], the Decision Tree classifier generalizes, in some cases, to the most frequent phone, that is the mid-high vowels [e,o]. The transcriber also had problems with the [k.s] (F1-score: 0.66) and [ʃ] (F1-score: 0.84). This result was also expected, both these phones are related to the grapheme <x> which, in Brazilian Portuguese, shows a very irregular behavior. In fact, <x> can be pronounced as [ʃ, s, z, k.s], depending on the word: "bruxa" *(witch)* [ʃ], "próximo" *(near)* [s]; "exame" *(test)* [z] and "axila" *(armpit)* [k.s].

4. Final Remarks

We presented the method we employed in building a pronunciation dictionary for Brazilian Portuguese. High F1-score values were achieved while transcribing most of the graphemes in Brazilian Portuguese and the dictionary can be considered robust enough for Large Vocabulary Continuous Speech Recognition (LVCSR) and Speech Synthesis. Although the rules we developed are language-specific, the architecture we used for compiling the dictionary, by using transcription rules and machine learning classifiers, can be successfully replicated in other languages. In addition, the entire dictionary, all scripts, algorithms and corpora were made publicly available.

5. Acknowledgements

Part of the results presented in this paper were obtained through research activity in the project titled "Semantic Processing of Brazilian Portuguese Texts", sponsored by *Samsung Eletrônica da Amazônia Ltda.* under the terms of Brazilian federal law number 8.248/91.

6. References

[1] R. Godwin-Jones, "Emerging technologies: Speech tools and technologies," *Language Learning and Technology*, vol. 13-3, pp. 4–11, 2009.

[2] F. Lewis, M. Gary and D. Charles, *Ethnologue: Languages of the World, Seventeenth edition*, ser. Seventeenth edition. Dallas, Texas: SIL International, 2013. [Online]. Available: http://www.ethnologue.com

[3] N. Neto, C. Patrick, A. Klautau, and I. Trancoso, "Free tools and resources for brazilian portuguese speech recognition," *Journal of the Brazilian Computer Society*, vol. 17, no. 1, pp. 53–68, 2011.

[4] R. I. Damper, Y. Marchand, M. Adamson, and K. Gustafson, "Comparative evaluation of letter-to-sound conversion techniques for english text-to-speech synthesis," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[5] T. Polyakova and A. Bonafonte, "Learning from errors in grapheme-to-phoneme conversion." in *INTERSPEECH*, 2006.

[6] A. Teixeira, C. Oliveira, and L. Moutinho, "On the use of machine learning and syllable information in european portuguese grapheme-phone conversion," in *Computational Processing of the Portuguese Language*. Springer, 2006, pp. 212–215.

[7] A. Veiga, S. Candeias, and F. Perdigão, "Developing a hybrid grapheme to phoneme converter for european portuguese," vol. 1, pp. 297–300, May 2013.

[8] H. Bussmann, G. Trauth, K. Kazzazi, and H. Bussmann, *Routledge dictionary of language and linguistics / Hadumod Bussmann ; translated and edited by Gregory Trauth and Kerstin Kazzazi*. Routledge, London ; New York :, 1996.

[9] S. Bergsma, P. McNamee, M. Bagdouri, C. Fink, and T. Wilson, "Language identification for creating language-specific twitter collections," in *Proceedings of the Second Workshop on Language in Social Media*, ser. LSM '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 65–74. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390374.2390382

[10] E. B. Bilcu and J. Astola, "A hybrid neural network for language identification from text," in *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*. IEEE, 2006, pp. 253–258.

[11] D. Trieschnigg, D. Hiemstra, M. Theune, F. de Jong, and T. Meder, "An exploration of language identification techniques for the dutch folktale database," in *Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage, LREC 2012*, P. Osenova, S. Piperidis, M. Slavcheva, and C. Vertan, Eds. Istanbul, Turkey: LREC organization, May 2012, pp. 47–51. [Online]. Available: http://doc.utwente.nl/82013/

[12] M. Zampieri, B. G. Gebre, and H. Nijmegen, "Automatic identification of language varieties: The case of portuguese," in *Proceedings of KONVENS*, 2012, pp. 233–237.

[13] T. C. Silva, *Fonética e fonologia do português: roteiro de estudos e guia de exercícios*. Contexto, 2005.

[14] C. Girelli, *Brazilian Portuguese Syllable Structure*. UMI, 1990. [Online]. Available: http://books.google.com.br/books?id=KRGMnQEACAAJ

[15] L. Bisol, "Vowel harmony: a variable rule in brazilian portuguese," *Language Variation and change*, vol. 1, pp. 185–198, 1989.

[16] A. Quicoli, "Harmony, lowering and nasalization in brazilian portuguese," *Lngua*, vol. 80, pp. 295–331, 1990.

[17] F. Delatorre and R. Koerich, "Production of epenthesis in ed-endings by brazilian efl learners," *Proceedings of the II Academic Forum*, p. 8, 2005.

[18] C. Oliveira, L. C. Moutinho, and A. J. S. Teixeira, "On european portuguese automatic syllabification." in *Proceedings of the Interspeech 2005*. ISCA, 2005, pp. 2933–2936. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2005/i05_2933.html

[19] V. Vasilévski, "Phonologic patterns of brazilian portuguese: a grapheme to phoneme converter based study," in *Proceedings of the Workshop on Computational Models of Language Acquisition and Loss*. Avignon, France: Association for Computational Linguistics, April 2012, pp. 51–60. [Online]. Available: http://www.aclweb.org/anthology/W12-0912

[20] W. Rocha and N. Neto, "Implementação de um separador silábico gratuito baseado em regras linguísticas para o português brasileiro," in *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 2013, pp. 108–115.

[21] Brasil, *Acordo ortogrfico da lngua portuguesa, de 14, 15 e 16 de dezembro de 1990*. Braslia, Brazil: Dirio do Congresso Nacional da Repblica Federativa do Brasil, Poder Executivo, 2009.

[22] Medialab, "Wikipedia extractor," http://medialab.di.unipi.it/wiki/Wikipedia_Extractor, 2013.

[23] I. Steinwart and A. Christmann, *Support vector machines*. Springer, 2008.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] I. Alves, *Neologismo: Criação lexical*, ser. Princípios (São Paulo). Editora Atica, 2001. [Online]. Available: http://books.google.com.br/books?id=7fIuAAAAYAAJ

[26] H. Weide, "The cmu pronouncing dictionary," 1998. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Evaluating phonetic spellers for user-generated content in Brazilian Portuguese

Omitted for blind review
-
-
-

No Institute Given

Abstract. Recently, spell checking (or spelling correction systems) has regained attention due to the need of normalizing user-generated content (UGC) on the web. UGC presents new challenges to spellers, as its register is much more informal and contains much more variability than traditional spelling correction systems can handle. This paper proposes two new approaches to deal with spelling correction of UGC in Brazilian Portuguese (BP), both of which take into account phonetic errors. The first approach is based on three phonetic modules running in a pipeline. The second one is based on machine learning, with soft decision making, and considers context-sensitive misspellings. We compared our methods with other methods on a human annotated UGC corpus of reviews of products. The machine learning approach surpassed all other methods, with 78.0% correction rate, very low false positive (0.7%) and false negative rate (21.9%).

1 Introduction

Spell checking is a very well-known and studied task of natural language processing (NLP), being present in applications used by the general public, including word processors and search engines. Most of the methods of spell checking are based on large dictionaries to detect non-words, mainly related to typographic errors caused by key adjacency or fast key stroking. Currently, with the recent boom of mobile devices, with small touchscreens and tiny keyboards, one can miss the keystrokes, hitting adjacent keys on the keyboard, thus spell checking has regained attention [1].

Dictionary-based approaches can be ineffective when the task is to detect and correct spelling mistakes which coincidentally corresponds to an existing word (real-word errors). Different from non-word errors, real-word errors are context dependent. Several approaches have been proposed to deal with these errors: mixed trigram models [2], confusion sets [3], improvements on the trigram-based noisy-channel model [4] and [5], use of GoogleWeb 1T 3-gram data set and a normalized and modified version of the Longest Common Subsequence string matching algorithm [6], a graph-based method using contextual and PoS features and the double metaphone algorithm to represent phonetic similarity [7]. As an example, although MS Word (from 2007 version to on) claims to include a contextual spelling checker, an independent evaluation of it found high precision but low recall in a sample of 1400 errors [8].

Errors due to phonetic similarity also impose difficulties to spell checkers. They occur when a writer knows well the pronunciation of a word but does not know how to spell it. This kind of error requires new approaches to combine phonetic models and models for correcting typographic and/or real-word errors. In [9], for example, the authors use a linear combination of two measures – the Levenshtein distance between two strings and the Levenshtein distance between the Soundex [10] code of two strings. In the last decade, some researchers have revisited spell checking issues motivated by web applications, such as search query engines and sentiment analysis tools based on natural language processing (NLP) of UGC, e.g. Twitter data or product reviews. Normalization of UGC has received great attention also because the performance of NLP tools (e.g. taggers, parsers and named entity recognizers) is greatly decreased when applied to UGC. Besides misspelled words, this kind of text presents a long list of problems, such as acronyms and proper names with inconsistent capitalization, abbreviations introduced by chat-speak style, slang terms mimicking the spoken language, loanwords in English as technical jargon, as well as problems related to ungrammatical language and lack of punctuation [11–14].

In [14] the authors propose a spell checker for Brazilian Portuguese (BP) to work on the top of Web text collectors. They have tested their method on news portals and on informal texts collected from Twitter in BP. However, they do not inform the error correction rate of the system. Furthermore, while their focus is on the response time of the application, they do not address real-word errors.

This paper presents two new spell checking methods for UGC in BP. The first of them deals with phonetic motivated errors, a recurrent problem in UCG not addressed by traditional spell checkers. The second one deals additionally with real-word errors. We present a comparison of these methods with a baseline system and JaSpell over a new and large benchmark corpus for this task. The corpus contains product reviews with 38,128 tokens and 4,083 annotated errors. Such corpus is also a contribution of our study[1]. This paper is structured as follows. In Section 2 we describe our methods, the setup of the experiments and the corpus we compiled. In Section 3 we present the results. In Section 4 we discuss related work on spelling correction of phonetic and real-word errors. To conclude, the final remarks are outlined in Section 5.

2 Experimental Settings and Methods

In this Section we present the four methods compared in our evaluation. Two of them are used by existing spellers, one is taken as baseline and the other is taken as benchmark. The remaining two are novel methods developed within the project reported herein. After describing in detail the novel methods, we present the corpus specifically developed to evaluate BP spellers, as well as the evaluation metrics.

[1] The small benchmark of 120 tokens used in [15] and [16] is not representative of our scenery.

2.1 Method I - Baseline

We use as a baseline the open source Java Spelling Checking Package, JaSpell[2]. JaSpell can be considered a strong baseline and it is employed at the tumba! Portuguese Web search engine to support interactive spelling checking of user queries. JaSpell classifies the candidates for a misspelled word according to the word frequency in a large corpus together with other heuristics, such as keyboard proximity or phonetic keys, provided by the Double Metaphone algorithm [17] for the English language. At the time this speller was delivered there was no version of these rules for the Portuguese language[3].

2.2 Method II - Benchmark

The method presented in [18] is taken as benchmark. It combines phonetic knowledge in the form of a set of rules and the algorithm Soundex. It was inspired by the analysis of errors of the same corpus of products' reviews [19] that inspired our proposals. Furthermore, as such method aims to be used for normalizing web texts, it performs automatic spelling correction. To increase the accuracy of the first hit, this method relies in some ranking heuristics. The strategies developed by the authors consider the phonetic proximity between the input wrong word and the candidates to substitute it. If the typed word does not belong to the lexicon, a set of candidates is generated by applying one and two edit distances from the original word and the words in the lexicon. Then a set of phonetic rules for Brazilian Portuguese codifies letters and digraphs which have similar sounds in a specific code. If necessary, the next step performs the algorithm Soundex, slightly modified for BP. Finally, if none of these phonetic-based algorithms is able to suggest a correction, the candidate with the minor edition-distance and which is the most frequent in a reference corpus is suggested. The lexicon used is the Unitex-PB[4] and the frequency list was taken from Corpus Brasileiro[5].

2.3 Method III - Grapheme-to-Phoneme based Method (GPM)

By testing the benchmark method, we noticed that many of the wrong corrections were related to a gap between the application of phonetic rules and the Soundex module. The letter-to-sound rules were developed specially for the spelling correction, therefore, they are very accurate for the task but have a low recall, since many words do not possess the misspelling patterns which they try to model. In contrast, the transcriptions generated by the adapted Soundex algorithm are too broad and many phonetically different words are given the same code. For instance, the words "perto" (*near*) and "forte" (*strong*) are both transcribed with the Soundex code "1630", in spite of being very distinct phonetically: "perto" corresponds to ['pɛh.tʊ], and "forte" to ['fɔh.tʃɪ].

[2] http://jaspell.sourceforge.net/

[3] Currently, a BP version of the phonetic rules can be found at http://sourceforge.net/projects/metaphoneptbr/

[4] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/

[5] http://corpusbrasileiro.pucsp.br/cb/

To fill this gap we propose the use of a general-purpose grapheme-to-phoneme converter to be executed prior to the Soundex module. We selected Aeiouado's grapheme-to-phoneme converter [20] for this purpose, since it consists of the state of the art in grapheme-to-phoneme transcription for Brazilian Portuguese. The usage of the grapheme-to-phoneme converter is a bit different from a simple pipeline. According to Toutanova [21], phonetic-based errors usually need larger edit distances to be detected. For instance, the word "durex" (*sellotape*) and one of its misspelled forms "duréquis" have an edit distance of 5 units, despite having exactly the same phonetic form: [duˈrɛ.kɪs]. Therefore, instead of simply increasing the edit distance, which would imply in having a larger number of candidates to filter, we decided to do the reverse process. We transcribed the Unitex-PB dictionary and stored it into a database, with the transcriptions as keys. Therefore, for obtaining words which are phonetic similar words, we transcribe the input word and look it up in the database. For instance, considering the "duréquis" example, we would first transcribe it as [duˈrɛ.kɪs], and then check if there are any words in the database with this transcription. In this case, it would return "durex", the expected form.

The only difference of GPM in comparison with Method II lies in the G2P transcription match, which takes place prior to Soundex. In spite of being better than the baseline because they tackle phonetic-motivated errors, Method II and GPM have a limitation: they do not correct real word errors. The following method is intended to overcome this shortcoming by using context information.

2.4 Method IV – GPM in a Machine Learning framework (GPM-ML)

Method IV has the advantage of bringing together many approaches to spelling correction into a machine learning framework. The architecture of the method is described in Figure 1.

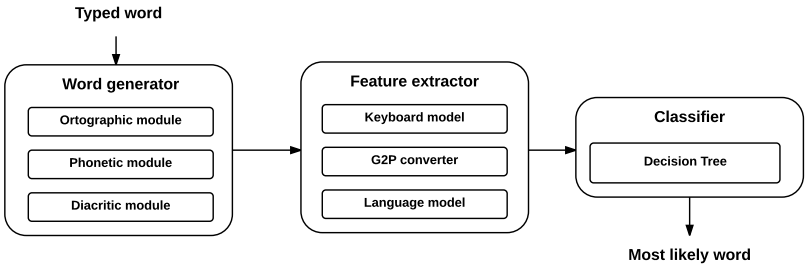


Fig. 1. *Architecture of the GPM-ML*

The method is based on three main steps: (i) word generation, (ii) feature extraction and (iii) word selection. The word generation phase encompasses three modules which produce a large number of suggestions, considering the following aspects: orthographic,

phonetic and diacritic similarities. For producing suggestions which are typographically similar, the Levenshtein distance is used. For each input word, we select all words in a dictionary which diverge from the input by at most 2 units. For instance, suppose the user intended to write "mesa" (*table*), but missed a keystroke and typed "meda" instead. The Levenshtein module would generate a number of suggestions including an edit distance of 1 or 2, such as "medo" (*fear*), "meta" (*goal*), "moda" (*fashion*), "nada" (*nothing*), "mexe" (*he/she moves*) etc. For computational efficiency, we stored the dictionary in a trie structure, in order to make it quickly searchable. A revised version of the Unitex-PB was employed as our reference dictionary (*circa* 550,000 words)[6].

As for phonetic similarity, the Aeiouado's grapheme-to-phoneme converter [20] was used to group phonetically related words. We transcribed the Unitex-PB word list phonetically and stored all word transcriptions along with their orthographic form into a database, exactly as we did for GPM. Thus for generating suggestions which are phonetically similar to the word typed by the user, we obtain its phonetic transcription and look it up in the database.

The diacritic module is responsible for generating words which are similar to the word typed by the user with respect to diacritic symbols. This module was proposed since we observed that most of the misspellings in the corpus were caused by a lack or misuse of diacritics. BP has five types of diacritics: accute (´), cedilla (¸), circumflex (ˆ), grave (`) and tilde (˜). The diacritics often indicate different vowel quality, timbre or stress. However, these symbols are rarely used in UGC, and the reader uses the the context to disambiguate the intended word. For allowing the speller to deal with this problem, the diacritic model generates, given a word input, all possible word combinations of diacritics. Once more, the Unitex-PB is used as reference.

After word generation, the feature extraction phase takes place. This phase is responsible for extracting relevant information from the list of words generated in the previous step. The aim is to allow the classifier to compare these words with the one typed by the user, in such way that the classifier is able to choose to keep the typed word or to substitute it by one of the generated suggestions.

As misspelling errors may be of different nature (such as typographical, phonological or related to diacritics), we try to select features that encompass all these phenomena. For each word suggestion produced in the word generation phase, we extract 14 features, as described in Table 1.

The probabilities come from a language model trained over a subset of the Corpus Brasileiro (*circa* 10 million tokens). Good-Turing smoothing is used to estimate the probability of unseen trigrams. After feature extraction, the word selection phase comes into play. It consists of a Decision Tree Classifier which was trained over the dataset presented in Section 2.5, with the features we discussed. The classifier was implemented through scikit-learn [22] and comprises an optimized version of the CART algorithm. Several other classification algorithms were tested, but since our features contain both nominal and ratio data, and since some of them are dependent, the Decision Tree Classifier achieved the best performance.

[6] The dictionary is available upon request.

Table 1. *List of features*

Feature	Description
1. TYPEDORGEN	whether the word was typed by the user or was produced in the word generation phase;
2. ISTYPO	1 if the word was generated by the typographical module; 0 otherwise;
3. ISPHONE	1 if the word was generated by the phonetic module; 0 otherwise;
4. ISDIAC	1 if the word was generated by the diacritic module; 0 otherwise;
5. TYPEDPROB	the unigram probability of the word typed;
6. GENUNIPROB	the unigram probability of the word suggestion;
7. TYPEDTRIPROB	the trigram probability of the word typed;
8. GENTRIPROB	the trigram probability of the word suggestion;
9. TYPOLEVDIST	the levenshtein distance between the typed word and the suggestion;
10. INSKEYDIST	the sum of the key insertion distances;
11. DELKEYDIST	the sum of the key deletion distances;
12. REPLKEYDIST	the sum of the key replacement distances;
13. KEYDISTS	the sum of all previous three types of key distances;
14. PHONELEVDIST	the levenshtein distance between of the phonetic transcription of the typed word and of the suggestion.

2.5 Dataset

The evaluation corpus was compiled especially for this research and is composed of a set of annotated product reviews, written by users on Buscapé[7], a Brazilian price comparison search engine. All misspelled words were marked, the correct expected form was suggested and the misspelling category was indicated. We used snowball sampling to obtain a reasonable amount of data with incorrect orthography. A list of ortographical errors with frequency greater than 3 in the the corpus of product reviews compiled by [19] was used to pre-select, from the same corpus, sentences with at least one incorrect word. Among those, 1,699 sentences were randomly selected to compose the corpus (38,128 tokens). All these sentences were annotated by two linguists with prior experience in corpus annotation.

The inter-rater agreement for the error detection task is described in Table 2.

Table 2. *Inter-rater agreement for the error detection task*

		Annot. B		
		Correct	**Wrong**	**Total**
Annot. A	**Correct**	33,988	512	34,500
	Wrong	76	3,559	3,635
	Total	34,064	4,071	38,135

The agreement was evaluated by means of the kappa test [23]. The κ value for the error detection task was 0.915 which stands for good reliability or almost perfect agreement [24]. The final version of the corpus used to evaluate all methods was achieved by submitting both annotations to an adjudication phase, in which all discrepancies were resolved. We noticed that most annotation problems consisted of whether or not to correct abbreviations, loanwords, proper nouns, internet slang, and technical jargon. In order to enrich the annotation and the evaluation procedure, we classified the misspellings into five categories:

[7] http://www.buscape.com.br/

1. TYPO: misspellings which encompass a typographical problem (character insertion, deletion, replacement or transposition), usually related to key adjacency or fast typing; e.g. ”obrsevei” instead of ”observei” (*I noticed*) and ”memso” instead of ”mesmo” (*same*).
2. PHONO: cognitive misspellings produced by lack of understanding of letter-to-sound correspondences, e.g. ”esselente” for ”excelente” (*excelent*), since both ”ss” and ”xc”, in this context, sound like [s].
3. DIAC: this class identifies misspellings which are related to the inserting, deleting or replacing diacritics in a given word, e.g. ”organizacao” instead of ”organização” (*organization*).
4. INT_SLANG: use of internet slang or emoticons, such as ”vc” instead of ”você” (*you*), ”kkkkkk” (to indicate laughter) or ”:-)”.
5. OTHER: other types of errors that do not belong to any of the above classes, such as abbreviations, loanwords, proper nouns, technical jargon; e.g. ”aprox” for ”aproximadamente” (*approximately*).

The distribution of each of these categories of errors can be found in Table 3. The difference between the total number of counts in Table 2 and 3 is caused by spurious orthographies which were reconsidered or removed in the adjudication phase. In addition to the five categories previously listed, we also classified the misspellings into either contextual or non-contextual; i.e. if the misspelled word corresponds to another existing word in the dictionary, it is considered a contextual error (or real-word error). For instance, if the intended word was “está” (*he/she/it is*), but the user typed “esta”, without the acute accent, it is classified as a contextual error, since “esta” is also a word in Brazilian Portuguese which means *this* FEM.

The corpus has been made publicly available[8] and intends to be a benchmark for future research in spelling correction for user generated content in BP.

Table 3. *Error distribution in corpus by category*

Misspelling type		Counts	% Total
TYPO	-	1,027	25.2
PHONO	Contextual	49	1.2
	Non-contextual	683	16.7
DIAC	Contextual	411	10.1
	Non-contextual	1,626	39.8
INT_SLANG	-	201	4.9
OTHER	-	86	2.1
Total/Avg		4,083	100.0

2.6 Evaluation Metrics

Four performance measures are used to evaluate the spellers. The *Detection rate* is the ratio between the number of errors detected and the total number of errors. The

[8] Link omitted for blind review.

Correction rate stands for the ratio between the number of corrected errors and the total number of errors. *False positive rate* is the ratio between the number of false positives (correct words that are wrongly detected as errors) and the total number of correct words. The *False negative rate* consists of the ratio between the number of false negatives (wrong words that are detected as correct) and the total number of errors. In addition, the correction hit rates are evaluated by misspelling categories. In the analysis, we do not take into account the "int_slang" and "other" categories, since both show a very irregular behavior and constitute specific types of spelling correction.

3 Discussion

In Table 4, we summarize all methods' results. As one can observe, the GPM-ML achieved the best overall performance, with the best results in at least three rates: detection, correction and false positive. Both methods we proposed in this paper, GPM and GPM-ML, performed better than the baseline in all metrics. However, GPM did not show any improvement in comparison to the benchmark. In fact, the addition of the grapheme-to-phoneme converter decreased the performance in what concerns to the correction rate. By analyzing the output of GPM, we noticed that there seem to be some overlapping information between the phonetic rules and the grapheme-to-phoneme module. Apparently, the phonetic rules were able to cover all cases which could be solved by adding the grapheme-to-phoneme converter. Therefore our hypothesis was not supported.

Table 4. *Comparison of the Methods*

	Rate			
Method	**Detection**	**Correction**	**FP**	**FN**
Baseline JaSpell	74.0%	44.7%	5.9%	26.0%
Benchmark Rules&Soundex	83.4%	68.6%	1.7%	**16.6%**
GPM	83.4%	68.2%	1.7%	**16.6%**
GPM-ML	**84.9%**	**78.1%**	**0.7%**	21.9%

All methods showed a low rate of false positives, the best value was found in GPM-ML (0.7%). The false positive rate is very important for spelling correction purposes and is related to the reliability of the speller. In the following we discuss the correction hit rates by misspelling categories. Table 5 presents a comparison among all methods.

The baseline JaSpell (Method I) presented an average correction rate of 44.7%. Its best results comprise non-contextual diacritic misspellings with a rate of 64.0%. Its worst result is found in contextual phonological errors, not a single case of this type of error was corrected by the speller. The typographical misspelling were also very troublesome for the baseline method, with a correction hit rate of 28.3%. These results indicate that the method is not suitable for real world applications which deal with user generated content. It is important to notice that the JaSpell was not developed specifically for this text domain, so its performance is much influenced by this fact.

The benchmark Rules&Soundex (Method II) achieved a correction rate of 68.6%, a relative gain of 53.4% in comparison to the baseline. The best results are, once more,

Table 5. *Comparison of Correction Rates*

Misspelling type	Errors	Correction rate by method			
		I	II	III	IV
Typo	1,027	28.3%	**56.3%**	53.0%	55.4%
Phono Contextual	49	0.0%	0.0%	0.0%	**8.1%**
Phono Non-contextual	683	48.2%	85.1%	**87.1%**	81.1%
Diac Contextual	411	9.2%	26.5%	26.5%	**64.5%**
Diac Non-contextual	1626	64.0%	82.2%	82.4%	**96.6%**
Total/Weighted Avg	3,796	44.7%	68.6%	68.1%	**78.0%**

related to the non-contextual diacritic misspellings (82.2%), which stand for the major class. The best improvements compared to the baseline appear in phonological errors that are influenced by context (85.1%), with a relative increase of 76.6%. These results are coherent with the results reported by [18], since they claim that the method focuses on phonetically motivated misspellings. As already mentioned, GPM (Method III) did not show any gain in comparison with the benchmark. As can be noticed, the grapheme-to-phoneme converter had a small positive impact in what regards to the phonological errors, raising the correction rate of non-contextual phonological misspellings from 85.1% to 87.1% (2.3% gain).

GPM-ML (Method IV) achieved the best performance among all methods in what regards to correction hit rate (78.0%). Some misspelling categories showed a very high correction rate, such as non-contextual diacritic errors (96.6%) and non-contextual phonological errors (81.1%). The trigram Language Model proved to be effective for capturing some contextual misspellings, as can be seen by the contextual diacritic correction rate (64.5%). However, the method was not able to properly infer contextual phonological misspellings (8.1%). We hypothesize that this result might be caused by the few number of contextual phonological instances in the corpus used for training (there were only 49 cases of contextual phonological misspellings). Such a small number of cases is not adequate for ensuring good performance by machine learning techniques. No significant improvement was found with respect to typographical errors (55.4%) in comparison to the other previous methods.

4 Related Work

The first approaches to spelling correction date back to Damerau[25] and address the problem by analyzing the edit distance of the words. He proposes a speller based on a reference dictionary and on an algorithm to check for out-of-vocabulary (OOV) words. The method assumes that words which are not found in the dictionary have at most one error, which was caused by a letter insertion, deletion, substitution or transposition. OOV words are then compared to the words from the dictionary. The one error threshold was established to avoid high computational cost. An improved error model for spelling correction, which works for letter sequences of lengths up to 5 and is also able to deal with phonetic errors was proposed by [26]. It embeds a noisy channel model for spell checking based on string to string edits. This model depends on the probabilistic modeling of sub-string transformations. As texts present several kinds of misspellings,

no single method will cover all of them, therefore it is very natural to combine methods which supplement each other. This approach was pursued by [21] who included information on pronunciation to the model of typographical errors correction. [21] and also [27] took the pronunciation of the misspelled words into account by using the technology of grapheme-to-phoneme converters. The later proposed the use of triphone analysis as a new correction strategy to combine phonemic transcription with trigram analysis, since they performed better than either grapheme-to-phoneme conversion or trigram analysis alone, in their evaluation. Our GPM method also combines models to correct typographical errors by using information on edition distance, information on pronunciation provided by a set of phonetic rules, on a grapheme-to-phoneme converter and finally on the output of the Soundex method. In this two-layer method these modules are put in sequence, as we take advantage of the high precision of the phonetic rules before trying the converter; typographical errors are corrected in the last pass of the process. We understand that the probabilistic classification framework used by [21] is very interesting and would provide better results to our two-layer method. Therefore, we decided to take advantage of a machine learning approach to decide how to correct a word, by using candidates generated by one and two edit distance, phonetic similarity and word combinations of diacritics. In our GPM-ML proposal, we adapted the output of a grapheme-to-phoneme converter which was developed for automatic speech recognition, and used it together with a keyboard model and a language model to provide features for a decision tree classifier. We had to broaden the transcriptions in order to deal with real-word errors related to diacritics, since the transcriptions are too much detailed for spelling correction purposes. With this new proposal one can deal with a special group of real-word errors caused by the presence or absence of diacritics, besides phonetic and typographic errors.

5 Final Remarks

We compared four spelling correction methods for UGC in BP, two of which consist of novel approaches and were proposed in this paper. The Method III (GPM) consisted of an upscale version of the benchmark method. In comparison to benchmark, it contained an additional module with a grapheme-to-phoneme converter. The grapheme-to-phoneme converter is intended to provide the speller with transcriptions that were not so fine-grained or specific as those generated by the phonetic rules and also not so coarse-grained as those created by Soundex. But our hypothesis was not supported. The Machine Learning version of GPM, the GPM-ML, however, presented a good overall, as it is the unique that addresses the problem of real word errors, and surpass all other methods in most situations. It reached 78.0% in correction rate, with very low false positive (0.7%) and false negative (21.9%), thus establishing as the new state of the art in spelling correction for UGC in BP. As for future work, we intend to improve GPM-ML by expanding the training database, by testing other language models as well as new phone conventions. In addition, we plan to more fully evaluate it into different testing corpora. We also envisage, in due course, the development of an internet slang module.

References

1. Duan, H., Hsu, B.P.: Online spelling correction for query completion. In: Proceedings of the 20th International Conference on World Wide Web. WWW '11, NY, USA, ACM (2011) 117–126
2. Fossati, D., Di Eugenio, B.: A mixed trigrams approach for context sensitive spell checking. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing. Volume 4394 of Lecture Notes in Computer Science., Springer (2007) 623–633
3. Fossati, D., Di Eugenio, B.: I saw tree trees in the park: How to correct real-word spelling mistakes. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC 2008. (2008) 897–901
4. Mays, E., Damerau, F.J., Mercer, R.L.: Context based spelling correction. Information Processing & Management **27**(5) (1991) 517–522
5. Wilcox-O'Hearn, A., Hirst, G., Budanitsky, A.: Real-word spelling correction with trigrams: A reconsideration of the mays, damerau, and mercer model. In: Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing. CICLing'08 (2008) 605–616
6. Islam, A., Inkpen, D.: Real-word spelling correction using google web 1tn-gram data set. In: In ACM International Conference on Information and Knowledge Management CIKM 2009. (2009) 1689–1692
7. Sonmez, C., Ozgur, A.: A graph-based approach for contextual text normalization. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing EMNLP 2014. (2014) 313 – 324
8. Hirst, G.: An evaluation of the contextual spelling checker of microsoft office word 2007 (2008)
9. Zampieri, M., Amorim, R.: Between sound and spelling: Combining phonetics and clustering algorithms to improve target word recovery. In: Proceedings of the 9th International Conference on Natural Language Processing PolTAL 2014. (2014) 438–449
10. Rusell, R.C.: US Patent 1261167 issued 1918-04-02. (1918)
11. Duran, M., Avanço, L., Aluísio, S., Pardo, T., Nunes, M.G.V.: Some issues on the normalization of a corpus of products reviews in portuguese. In: Proceedings of the 9th Web as Corpus Workshop WaC-9, Gothenburg, Sweden (April 2014) 22–28
12. De Clercq, O., Schulz, S., Desmet, B., Lefever, E., Hoste, V.: Normalization of dutch user-generated content. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013. (2013) 179–188
13. Han, B., Cook, P., Baldwin, T.: Lexical normalization for social media text. ACM Trans. Intelligent System Technology **4**(1) (February 2013) 5:1–5:27
14. Andrade, G., Teixeira, F., Xavier, C., Oliveira, R., Rocha, L., Evsukoff, A.: Hasch: High performance automatic spell checker for portuguese texts from the web. Procedia Computer Science **9**(0) (2012) 403 – 411
15. Martins, B., Silva, M.J.: Spelling correction for search engine queries. In: Proceedings of the 4th International Conference EsTAL 2004 España for Natural Language Processing. Volume 3230 of Lecture Notes in Computer Science. (2004) 372–383
16. Ahmed, F., Luca, E.W.D., Nürnberger, A.: Revised n-gram based automatic spelling correction tool to improve retrieval effectiveness. Polibits (12 2009) 39–48
17. Philips, L.: The double metaphone search algorithm. C/C++ Users Journal **18**(6) (2000)
18. Avanço, L., Duran, M., Nunes, M.G.V.: Towards a phonetic brazilian portuguese spell checker. In: Proceedings of ToRPorEsp Workshop PROPOR 2014, São Carlos, Brazil (2014) 24–31

19. Hartmann, N., Avanço, L., Balage, P., Duran, M., Nunes, M.G.V., Pardo, T., Alusio, S.: A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation LREC'14. (2014) 3866–3871
20. Mendonça, G., Aluísio, S.: Using a hybrid approach to build a pronunciation dictionary for brazilian portuguese. In: Proceedings of the 15th Annual Conference of the International Speech Communication Association INTERSPEECH 2014, Singapure (2014)
21. Toutanova, K., Moore, R.C.: Pronunciation modeling for improved spelling correction. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02 (2002) 144–151
22. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12** (2011) 2825–2830
23. Carletta, J.: Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics **22**(2) (June 1996) 249–254
24. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1) (1977) pp. 159–174
25. Damerau, F.J.: A technique for computer detection and correction of spelling errors. Communications of ACM **7**(3) (mar 1964) 171–176
26. Brill, E., Moore, R.C.: An improved error model for noisy channel spelling correction. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. ACL '00 (2000) 286–293
27. van Berkel, B., Smedt, K.D.: Triphone analysis: A combined method for the correction of orthographical and typographical errors. In: Proceedings of the Second Conference on Applied Natural Language Processing, Austin, Texas, USA (February 1988) 77–83

A Method for the Extraction of Phonetically-Rich Triphone Sentences

Gustavo Mendonça*, Sara Candeias†‡, Fernando Perdigão†, Christopher Shulby*,
Rean Toniazzo§, Aldebaro Klautau¶ and Sandra Aluísio*
*Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo – São Carlos, Brazil.
†Instituto de Telecomunicações,
Universidade de Coimbra – Coimbra, Portugal.
‡Microsoft Language Development Center – Lisbon, Portugal.
§Departamento de Engenharia de Materiais,
Universidade Federal de São Carlos – São Carlos, Brazil.
¶Laboratório de Processamento de Sinais,
Universidade Federal do Pará – Belém, Brazil.
Email: gustavom@icmc.usp.br, saracandeias@co.it.pt, fp@co.it.pt, chrisshulby@gmail.com,
reantoniazzo@gmail.com, a.klautau@ieee.org, sandra@icmc.usp.br

***Abstract*—A method is proposed for compiling a corpus of phonetically-rich triphone sentences; i.e., sentences with a high variety of triphones, distributed in a uniform fashion. Such a corpus is of interest for a wide range of contexts, from automatic speech recognition to speech therapy. We evaluated this method by building phonetically-rich corpora for Brazilian Portuguese. The data employed comes from Wikipedia's dumps, which were converted into plain text, segmented and phonetically transcribed. The method consists of comparing the distance between the triphone distribution of the available sentences to an ideal uniform distribution, with equiprobable triphones. A greedy algorithm was implemented to recognize and evaluate the distance among sentences. A heuristic metric is proposed for pre-selecting sentences for the algorithm, in order to quicken its execution. The results show that, by applying the proposed metric, one can build corpora with more uniform triphone distributions.**

I. Introduction

In what regards to speech technology, although there are some studies which employ words [1], syllables [2] and monophones [3] to develop Automatic Speech Recognition (ASR) and Text to Speech (TTS) systems, most of the current research widely makes use of contextual phone units, such as triphones and diphones.

The issue of developing a phonetically-rich triphone sentences corpus is of great significance for many areas of knowledge. In many applications of ASR and speech synthesis, for instance, rich speech databases are important for properly estimating the acoustic models [4]. In speech therapy, phonetically-rich sentences are often employed in reading aloud tasks so as to assess the speech production of patients in various phonetic/phonological contexts [5]. Laboratory phonologists are also interested in such corpora in order to develop prompts for analyzing speech production and variability [6].

Formally, the task discussed in this work can be described as follows: given a corpus K with s sentences, find a subset P containing s_p sentences, such that the triphones that compose s_p holds a uniform distribution as much as possible. Despite its apparent simplicity, in what concerns to computational complexity, the task cannot be considered a simple one. Since it has a combinatorial nature, it lacks a polynomial-time solution and should be regarded as an intractable problem [7].

We evaluate the proposed method in building a phonetically-rich triphone sentences corpus for Brazilian Portuguese. The sentences come from the Portuguese Wikipedia dump [8], which was converted into plain text, segmented and phonetically transcribed. The algorithm employs a greedy approach to select sentences, in a way such that the triphone distribution in the selected sentences is as uniform as possible. In order to expedite its execution, a heuristic metric is proposed to pre-select sentences for the algorithm, favoring the least frequent triphones over the most frequent ones.

The remainder of this paper is organized as follows. In Section II, we briefly describe the related work available in the literature. In Section III, we describe the method proposed. In Section IV, we evaluate it by building a phonetically-rich corpus for Brazilian Portuguese. The final remarks are outlined in Section V.

II. Related Work

Speech can be analyzed in a myriad of forms. The phonetic or phonological structure of a language can be described through phones, phonemes, syllables, diphones, triphones, feet, etc. For languages such as Mandarim, in which tones have a phonological value, one must even posit units such as tonemes in order to properly describe speech phenomena [9].

Many methods have been proposed for extracting phonetically-balanced corpora, that is to say corpora made of sentences which reproduce the triphone distribution of a given language [10][11][12][13].

It is known that many linguistic phenomena, including triphone sets, show a Zipfian distribution [14]. A phonetically-balanced corpus, for this reason, is a corpus which follows

Zipf's law in representing each triphone inversely proportional to its rank in the frequency table. These kinds of corpora are important specially for Large Vocabulary Continuous Speech Recognition (LVCSR), where unbalanced triphone representations can achieve better Word Error Rates (WER). However, phonetically-balanced corpora are not adequate for many other tasks, even regarding speech recognition. When building a system to assess one's pronunciation quality or to synthesize speech, for instance, more accurate results can be attained by using uniform triphone representations, i.e. phonetically-rich corpora.

Phonetically-rich corpora in our work are those which show sentences with a high variety of triphones, distributed in a uniform fashion regardless their representation in the language. In other words, in order to build such corpora, Zipf's law must be nullified, by favoring less frequent triphones and disfavoring more frequent ones. However, there are studies that consider other definitions and even other basic units to build phonetically-rich corpora.

In Abushariah et al. [10], the concept of "rich" is used in the sense that the set must contain all the phonemes of Arabic language (the chosen language for their study) but without a need for a uniform distribution. The set of sentences was handmade developed by linguists/experts. They used a set of 663 words, also defined by hand, and then Arabic independent sentences have been written using the 663 phonetically-rich words. The final database consists of 367 sentences with 2 to 9 words per sentence.

Arora et al. [15] considered syllables as the basic unit to extract, in an automatic way, phonetically-rich sentences from a large text corpus from Indian languages, justifying their choice because a syllable is the smallest segment of the uterance. In their process to extract the sentences for a given corpus, the chosen set should have the same distribution of syllabic words and also the same distribution of consonant, vowel and other symbols.

Nicodem et al. [16] deals specifically with Brazilian Portuguese and proposed a method based on genetic algorithms to select a set of sentences for a speech syntesis system. Their goal was to select a recording corpus that would improve the phonetic and prosodic variability of the system. They tried to fulfill the gap of phonetically-balanced corpora available for Brazilian Portuguese, since the available corpora disregards prosodic features. They evaluated it through the CETENFolha corpus (www.linguateca.pt/cetenfolha/) which has circa 1,5 million sentences in order to gather 4,000 sentences phonetically- and prosodically- rich. Their approach is composed of 4 stages, including grapheme-to-phoneme conversion, prosodic annotation, feature vector representation, and selection. The authors obtained prosodic features based on the pitch, therefore identifying tone events for each syllable (N, H+, H-, H, L, and L-, where H and L stands for high and low, respectively, and N for neutral). Using these features to represent each sentence, they developed a genetic algorithm (GA) to select a subset. Their paper, however, does not discuss how the GA fitness function meets both constraints (phonetic and prosodic).

III. Method

A. Unit of analysis

Contextual phone units are extensively applied to speech technology systems given their ability to encompass allophonic variation and coarticulation effects, specially triphones. A triphone is represented as a sequence $(p_{left}$ - p - $p_{right})$, where p_{left} is the phone which precedes p and p_{right} is the one which follows it. Table I presents a comparison of the word *speech* transcribed using monophones and triphones.

Word	Monophone Form	Triphone Form
speech	[s p i tʃ]	[#-s-p s-p-i p-i-tʃ i-tʃ-#]

TABLE I. A comparison between monophone and triphone transcription.

As one might observe, triphones are capable of describing the surronding environment of a given phone and this has a huge impact in the performance of acoustic models for speech recognition or speech synthesis. Given the above reasons, we chose triphones as the unit of analysis for our algorithm.

B. Heuristic Metric

For the expedition of the sentence extraction through the greedy algorithm, due to its high time complexity order, we set a heuristic metric to pre-select sentences and rank them according to the triphones they contained. The metric uses the probability of the triphones in the corpus in order to favor the least frequent triphones over the most frequent ones. It consists of a summation of the reciprocal probability for each triphone in the sentence.

Formally, this can be defined in the following way. Consider a corpus K consisting of a set of sentences $S = \{s_1, s_2, s_3, ..., s_n\}$. Each sentence s is formed by m triphones, represented as $T = \{t_1, t_2, t_3, ..., t_m\}$. The *a priori* probability of the triphones can be calculated straightforwadly: let $P_K(t_i)$ be the probability of the triphone t_i in the corpus K, then $P_K(t_i)$ is the number of times t_i occur divided by the total number of triphones in K. For that matter, a sentence s can be considered phonetically-rich if it possess many triphones with low probability of occurence. Therefore, we define the phonetic richness of a sentence s as the summation of its triphones' reciprocal probabilities:

$$\varrho(s) = \sum_{i=1}^{m} \frac{1}{P_K(t_i)} \quad (1)$$

C. Algorithm

Our algorithm for extracting rich sentences was implemented in Python and follows a greedy strategy. The distance metric is calculated through the SciPy library [17].

Greedy algorithms have been widely used in Computer Science, when the optimum solution of the problem can not be guaranteed [18]. Greedy strategies make locally optimal choices hoping to find the global optimum. Notwithstanding, in many cases, greedy algorithms have been notorious for jams at local maxima, since the best solution for a given problem may not concur with the sum of each partial best choice.

However, for the extraction of phonetically rich sentences, this approach is suitable, owing to the fact that it is computationally intractable to analyze all possible sets of sentences.

We initialize the algorithm by applying the heuristic metric described in Section III-B to all sentences in the corpus. After this, all sentences are ranked in descending order and the first 50,000 sentences with the best values are selected. This metric was proposed because the algorithm has an order of $O(mn^2)$ time complexity, where n is the number of sentences and m the number of selected triphones, and its execution was slow considering all the sentences available in the corpus. Afterwards, the algorithm loops through 50,000 sentences and calculates the euclidean distance between the triphone distribution of the set made up with the selected sentences and the current sentence to an ideal corpus, containing equiprobable triphones. The sentence with the minimum value is appended to a list of selected sentences and removed from the corpus. Then the loop starts over, considering for the calculation of the distance not just each sentence in isolation, but a set comprising each remaining sentence in the corpus together with the sentences already selected in the last step. When the list reaches n selected sentences, the execution is suspended. The pseudocode for the algorithm is described below.

```
Corpus <- List of available sentences
Selected <- [] // List of selected sentences
Metrics <- [] //List made of tuples with sentences
              and euclidean distance values
Ideal <- Ideal corpus, with all equiprobable triphones

while length(Selected) < n do:
  for Sentence in Corpus:
    calculate distance between Sentence+Selected and Ideal
    append Sentence and its metric in the list Metrics
  BestSentence <- select the sentence in the loop with the
                  minimum distance
  append BestSentece to Selected
  clear the Metrics list
end.
```

IV. Example Evaluation

A. Corpus

As a proof-of-concept we evaluated our method by building a phoneticaly-rich corpus for Brazilian Portuguese. The original database of sentences consisted of the Wikipedia dump produced on 23rd January 2014. Table II summarizes the data.

Articles	Word Tokens	Word Types
~820,000	168,823,100	9,688,039

TABLE II. Portuguese Wikipedia Summary – Dumped on 23rd January 2014.

In order to obtain only plain text from Wikipedia articles, we used the software WikiExtractor [19], to strip all of the MediaWiki markups and other metadata. Then, we segmented the output into sentences, by applying the Punkt sentence tokenizer [20]. Punkt is a language-independent tool, which can be trained to tokenize sentences. It is distributed together with NLTK [21], where it already comes with a model for Portuguese, trained on the Floresta Sintá(c)tica Treebank [22].

Following, each sentence was transcribed phonetically by using a pronunciation dictionary for each language variety. We employed the UFPAdic 3.0 [23], developed for Brazilian Portuguese, which contains 38 phones and 64,847 entries. Triphones were generated dynamically, based on the transcription registered in the dictionary. Cross-word triphones were considered in the analysis along with cross-word short pause models. Given its encyclopedic nature, many sentences in Wikipedia present dates, periods, percentages and other numerical information. For this reason, we decided to supplement the dictionary, by introducing the pronunciation of numbers from 0 to 2014. The pronunciations were defined manually and embedded into the dictionary. The transcription task was carried out in the following way: a Python script was developed to loop over each sentence and check if all its belonging words were listed on the dictionary. If all the words were listed, the sentence was accepted, otherwise rejected. Due to the fact that many words which occur in Wikipedia were not registered in the pronunciation dictionary, a large number of sentences had to be discarded. Details are described in Table III.

Total Sentences	Used	Used/total
7,809,647	1,229,422	15.7%

TABLE III. Sentences' summary after WikiExtractor and Punkt.

Some pilot experiments showed that the metric benefited sentences which were too long, as they had more triphones; or too short, as some of them had very rare triphones. The problem with long sentences is that they can be too complex for a recording prompt, inducing speech disfluencies such as pauses, false starts, lenghtenings, repetitions and self-correction [24]. In addition, the short sentences selected by the algorithm were usually only nominal, containing titles, topics or proper names; therefore, they would not be adequate for sentence prompts. For this reason, we filtered the sentences, selecting only those which had an average size (i.e. between 20 and 60 triphones, and more than four words). Further information is given in Table IV. After that, we applied the heuristic metric described in Section IV-A, and the top 50,000 sentences were selected (= 2,340,237 triphone tokens and 10,237 triphone types).

Total Sentences	Short	Average	Long
1,229,422	15,581	873,546	340,295

TABLE IV. Sentences' summary after the length filter.

B. Discussion

For this example evaluation, we discuss the extraction of 250 phonetically-rich sentences. Table V describes some triphone statistics for different sets of sentences extracted with the method proposed. The first column presents the number of extracted sentences; the second number of different triphones or triphone types; the third the number of triphone tokens; and the last the triphone type/token ratio which can be used to measure the method's performance. Owing to the fact that no other methods for the extraction of phonetically-rich triphone sentences were found in the literature, we established a list of random sentences as the baseline for comparison. Table VI contains the data regarding sentences selected randomly. The list of random sentences derives from the pool of 50,000 sentences described in Section IV-A. Ten different seed states were used in order to ensure randomness, the average of these results are presented.

Sentences	Triphone Types	Triphone Tokens	Type/Token
25	923	928	0.99
50	1485	1541	0.96
75	1965	2151	0.91
100	2389	2774	0.86
125	2736	3384	0.81
150	3091	4075	0.76
175	3390	4736	0.72
200	3715	5477	0.68
225	3991	6200	0.64
250	4189	6908	0.61

TABLE V. TRIPHONE RESULTS FROM THE EXTRACTION OF SENTENCES THROUGH OUR METHOD.

Sentences	Triphone Types	Triphone Tokens	Type/Token Ratio
25	774	1121	0.69
50	1318	2037	0.65
75	1713	3093	0.55
100	1917	3968	0.48
125	2352	5166	0.46
150	2564	6110	0.42
175	2820	7375	0.38
200	2961	8000	0.37
225	3211	9578	0.34
250	3335	10482	0.32

TABLE VI. TRIPHONE RESULTS FROM THE SENTENCES TAKEN RANDOMLY.

As it can be seen through the type/token triphone ratio, the method is capable of extracting sentences in a much more uniform way. For 250 sentences, our method was capable of extracting 4189 distinct triphones (40,9% of all types in the corpus), as opposed to 3335 (32,5%) in the random set; a difference of 854 novel distinct triphones. Furthermore, this higher number of distinct triphones was achieved with less triphone tokens (6908 *vs.* 10482), in a way that the type/token ratio for the method we propose was almost double the baseline: 0.61 in contrast to 0.32. Considering sets with different numbers of sentences, the method outperformed the random selection in all experiments. A Kolmogorov-Smirnov Test (K–S Test) confirms that the sentences selected through our method are closer to a uniform distribution than the ones extracted randomly.

One can observe that, as the number of selected sentences increases, the type/token ratio decreases. It may be the case that, after a huge number of sentences, the method's output converges to a limit such that no statistical significance can be noticed while comparing to a random selection. However, given time limitations, it was not feasible to analyze such a situation. As the number of selected sentences increases so does the number of triphones for comparison. After a while, the number of triphones for comparison becomes so large that the algorithm's execution time might not be proper for practical applications.

Additionally, the algorithm's output needs to be revised. Despite all our caution in the data preparation process, we noticed that some of the sentences selected by the algorithm were, in fact, caused by mistakes from the pronunciation dictionary. Foreign and loan words are known to be a problem for grapheme to phoneme conversion because they do not follow the orthographic patterns of the target language [25]. Several sentences selected by our algorithm contained foreign words which were registered in the dictionary with abnormal pronunciations, such as *Springsteen* [sprĩgsteẽ], *hill* [iww], *world* [wohwdʒ]. Since no other words are registered with the triphones [e-e+ẽ] or [e-ẽ+#] except for *Springsteen*, the algorithm ends up by selecting the sentence in which it occurs. Seeing that our method of comparing triphone distributions is greedy, our algorithm is fooled into believing that these are rare jewels. While this may be the case either way, the algorithm cannot function properly with incorrect transcriptions. A corpus with 100 revised sentences extracted by this method can be found in the Appendix.

V. FINAL REMARKS

We proposed a method for compiling a corpus of phonetically-rich triphone sentences. It was evaluated for Brazilian Portuguese. All sentences considered come from the Portuguese Wikipedia dumps, which were converted into plain text, segmented and transcribed. Our method consisted of comparing the distance between the triphone distribution of the sentences to a uniform distribution, with equiprobable triphones. The algorithm followed a greedy strategy in evaluating the distance metric. The results showed that our method is capable of extracting sentences in a much more uniform way, while comparing to a random selection. For 250 sentences, we were able to extract 854 new distinct triphones, in a set of sentences with a much higher type/token ratio. However, the method has its limitations. As discussed, it depends entirely on the quality of the pronunciation dictionary. If the pronunciation dictionary has some incorrect words, it might be the case that the algorithm favors them, if they possess triphone types not registered in other words. As a future work, we intend to define a method that recognizes foreign words and excludes them from the selected sentences. We also plan in applying the method to others corpora, e.g. CETENFolha, in order to make the results comparable with other studies for Brazilian Portuguese, such as Nicodem et al. [16]. All resources developed in this paper are freely available on the web[1].

REFERENCES

[1] R. Thangarajan, A. M. Natarajan, and M. Selvam, "Word and triphone based approaches in continuous speech recognition for Tamil language," *WSEAS Trans. Sig. Proc.*, vol. 4, no. 3, pp. 76–85, 2008.

[2] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, vol. 9, pp. 358–366, 2001.

[3] A. Kumar, M. Dua, and T. Choudhary, "Article: Continuous Hindi speech recognition using monophone based acoustic modeling," *IJCA Proceedings on International Conference on Advances in Computer Engineering and Applications*, vol. ICACEA, no. 1, pp. 15–19, March 2014.

[4] L. Rabiner and R. Schafer, "Introduction to digital speech processing," *Foundations and trends in signal processing*, vol. 1, pp. 1–194, 2007.

[5] A. P. Mendes, A. N. d. Costa, A. D. Martins, A. F. O. Fernandes, S. M. D. d. R. Vicente, and T. C. S. Freitas, "Contributos para a construção de um texto foneticamente equilibrado para o Português-Europeu," *Revista CEFAC*, vol. 14, pp. 910–917, 10 2012.

[6] J. B. P. Pierrehumbert, M. E. Beckman, and D. R. Ladd, "Conceptual foundations of phonology as a laboratory science," in *Phonological knowledge: Conceptual and empirical issues*. Oxford University Press., 2000, pp. 273–304.

[7] R. Sedgewick and P. Flajolet, *An introduction to the analysis of algorithms*. Addison-Wesley-Longman, 2013.

[8] Wikimedia, "Portuguese Wikipedia database dump backup," http://dumps.wikimedia.org/ptwiki/20140123/, 2014.

[1] http://nilc.icmc.usp.br/listener

[9] X. Lei, M. yuh Hwang, and M. Ostendorf, "Incorporating tone-related MLP posteriors in the feature representation for Mandarin ASR," in *In Proc. Eur. Conf. Speech Communication Technology*, 2005, pp. 2981–2984.

[10] M. A. M. Abushariah, R. N. Ainon, R. Zainuddin, M. Elshafei, and O. O. Khalifa, "Phonetically rich and balanced text and speech corpora for Arabic language," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 601–634, 2012.

[11] J.-L. Shen, H.-M. Wang, R.-Y. Lyu, and L.-S. Lee, "Incremental speaker adaptation using phonetically balanced training sentences for Mandarin syllable recognition based on segmental probability models." in *ICSLP*. ISCA, 1994. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/icslp1994.html#ShenWLL94

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus cdrom," 1993.

[13] E. Uraga and C. Gamboa, "VOXMEX speech database: Design of a phonetically balanced corpus," in *LREC*. European Language Resources Association, 2004.

[14] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[15] K. Arora, S. Arora, K. Verma, and S. S. Agrawal, "Automatic extraction of phonetically rich sentences from large text corpus of Indian languages." *INTERSPEECH*, 2004.

[16] M. Nicodem, I. Seara, R. Seara, D. Anjos, and R. Seara-Jr, "Seleção automática de corpus de texto para sistemas de síntese de fala," *XXV Simpósio Brasileiro de Telecomunicações - SBrT 2007*, 2007.

[17] E. Jones, T. Oliphant, and P. Peterson, "SciPy: Open source scientific tools for Python," http://www.scipy.org/, 2014.

[18] B. Coppin, "Inteligência artificial," *Rio de Janeiro: LTC*, 2010.

[19] Medialab, "Wikipedia extractor," http://medialab.di.unipi.it/wiki, 2013.

[20] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection." *Computational Linguistics*, vol. 32, no. 4, pp. 485–525, 2006.

[21] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

[22] C. Freitas, P. Rocha, and E. Bick, "Floresta sintá(c)tica: Bigger, thicker and easier," in *Computational Processing of the Portuguese Language*. Springer, 2008, pp. 216–219.

[23] N. Neto, C. Patrick, A. Klautau, and I. Trancoso, "Free tools and resources for Brazilian Portuguese speech recognition," *Journal of the Brazilian Computer Society*, vol. 17, no. 1, pp. 53–68, 2011.

[24] M. Watanabe and R. Rose, "Pausology and hesitation phenomena in second language acquisition," *The Routledge Encyclopedia of Second Language Acquisition*, pp. 480–483, 2012.

[25] J. Steigner and M. Schrder, "Cross-language phonemisation in German text-to-speech synthesis." in *INTERSPEECH 2007*. ISCA, 2007, pp. 1913–1916.

APPENDIX: EXAMPLES OF THE EXTRACTED SENTENCES

Number of Sentences: 100; **Number of Triphone Types:** 2307; **Number of Triphone Tokens:** 2959; **Type/Token Ratio:** 0.78.

1) A ilha fica tão próxima da praia que, quando a maré baixa, pode ser atingida a pé.
2) Diadorim é Reinaldo, filho do grande chefe Joca Ramiro, traído por Hermógenes.
3) A Sicília tem alguns moinhos ainda em bom estado de conservação que lhe dão beleza e encanto.
4) Em geral, chegaram ao Brasil como escravos vindos de Angola, Congo, e Moçambique.
5) A sardinha é um peixe comum nas águas do mar Mediterrâneo.
6) Possuem esse nome pois costumam viver na plumagem dos pombos urbanos.
7) É brilhante, doce e muito harmônico, sem presença de metal na voz.
8) Para fechar Alessandro Del Piero fez outro aos 121'.
9) Roman Polanski dirige Chinatown com Jack Nicholson.
10) A atriz sabe falar fluentemente espanhol.
11) Eles achavam Getúlio Vargas um problema.
12) Oppenheimer captura cavalo com peão.
13) Um bago tem tamanho médio não uniforme.
14) Segundo relatório da força aérea belga há confrontos com a União Soviética.
15) É irmão do também antropólogo Gilberto Velho.
16) Ganhou sete Oscar e oito Emmy.
17) Qual é minha perspectiva agora?
18) Ela é um fantasma verde, feminino!
19) Justin em seguida volta no tempo.
20) Nós fizemos um álbum do Korn.
21) Desde então Edílson é fã dessas bandas.
22) Há um só senhor uma só fé um só batismo.
23) Ivan Lins faria um show em Mossoró à noite.
24) Cresceram maior que um gato.
25) Há locações disponíveis em Tóquio no Japão.
26) Preso a um tronco nenhum lugar é seguro!
27) Hoje é professor emérito da UFBA.
28) Veio até aqui e não vai mergulhar?
29) Luís Jerônimo é um jovem rico.
30) Na hora pensei: "tenho que fazer isso?"
31) A campanha teve coordenação de Sanches.
32) A mulher que você me deu, fugiu.
33) Eu nunca tive um encontro com Bianca.
34) Homer jura vingança a Burns.
35) Beijo, me liga e amanhã sei lá!
36) Um colégio é como um ser vivo.
37) Sophie é filha de um amigo gay de Alan Greg.
38) Xuxa guarda rancor e é ambiciosa.
39) No mesmo ano conhece Aldir Blanc em Viena.
40) É um imenso painel reunindo um elenco famoso.
41) A Sé integra três belos órgãos.
42) Em ambos, Shannon conquistou medalha.
43) A terra é abundante em recursos como vinagre e óleo vegetal.
44) Faça sua escolha e bom jogo!
45) Quem é que poderia sonhar com algo assim?
46) Ela é ruiva com olhos azuis.
47) Deu a louca na Chapeuzinho!
48) De onde venho e para onde vou?
49) Eu choro e sofro tormentas!
50) Um falcão pousa em um pedregulho.
51) Ninguém tenha medo, nem fraqueza!
52) É membro do grupo Monty Python.
53) A sondagem de Senna pela Benetton e a chegada à kart.
54) Isto é um negócio e a única coisa que importa é ganhar.
55) Robert é um forte glutão da equipe.
56) Um bárbaro no exército romano?
57) Infância e juventude em Linz.
58) Já ir à Argentina era muito bom!
59) Fiquei com inveja dele.
60) Há dragões ao redor do mundo!
61) Edmond é pai do biólogo Jean.
62) A mãe lhe telefonava às vezes.
63) Tonho é tímido, humilde e sincero.
64) André Jung ocupa um lugar central no fórum.
65) Lois pergunta: "você é um homem ou um alienígena?"
66) Sua voz é um assobio fino e longo.
67) Por isso é sempre bom conferir!
68) Celso Lafer recuperou a jóia e devolveu-lhe.
69) É próxima ao Rio Parnaíba.
70) Lendo aquilo fica bem difícil.
71) A faculdade de John Oxford até hoje possui fãs fiéis.
72) Existe uma crença moderna no dragão chinês.
73) Sean Connery já sugeriu que Gibson fosse James Bond.
74) A raiz dos dentes é longa.
75) Essa noite produziu um feito singular.
76) Fim da Segunda Guerra Mundial.
77) –No Zorra, eu fazia humor rasgado.
78) Charles vê um homem ser morto em um tiroteio.
79) Tinham um novo senhor agora.
80) É comum ocorrerem fenômenos ópticos com estas nuvens.
81) Era um cão de pelo escuro e olhos negros.
82) Há títulos na região tcheca da Tchecoslováquia.
83) Raquel Torres vai investigar a área.
84) Clay foge e leva a jovem Jane como refém.
85) Djavan jogou futebol e hóquei no gelo na infância.
86) A origem do fagote é bastante remota.
87) Um jedi nunca usa a força para lucro ou ganho pessoal.
88) Chamavam José Alencar de Zezé.
89) Um código fonte é um sistema complexo.
90) A igreja tem um altar barroco.
91) Luís Eduardo pronunciou a senha: "esgoto".
92) Quanto ao sexo: macho ou fêmea?
93) A rádio Caxias cumpriu esse papel.
94) Roger Lion é um campeão orgulhoso que ama boxe.
95) Um outeiro é menor que um morro.
96) Hitoshi Sakimoto nasceu em Yokohama.
97) Nenhum isótopo do urânio é estável.
98) Chicago é um bairro tranquilo e festivo.
99) Hong Kong continua a utilizar a lei comum inglesa.
100) Só cinco funcionam como museus.

A prototype system for automatic speech recognition and evaluation of Brazilian-accented English

Gustavo Mendonça[1], **Sandra Aluisio**[1]

[1]Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo (ICMC/USP)

gustavoauma@gmail.com, sandram@icmc.usp.br

Abstract. *This meta-paper describes the style to be used in articles and short papers for SBC conferences. For papers in English, you should add just an abstract while for the papers in Portuguese, we also ask for an abstract in Portuguese ("resumo"). In both cases, abstracts should not have more than 10 lines and must be in the first page of the paper.*

Resumo. *Este meta-artigo descreve o estilo a ser usado na confecção de artigos e resumos de artigos para publicação nos anais das conferências organizadas pela SBC. É solicitada a escrita de resumo e abstract apenas para os artigos escritos em português. Artigos em inglês deverão apresentar apenas abstract. Nos dois casos, o autor deve tomar cuidado para que o resumo (e o abstract) não ultrapassem 10 linhas cada, sendo que ambos devem estar na primeira página do artigo.*

1. General Information

All full papers and posters (short papers) submitted to some SBC conference, including any supporting documents, should be written in English or in Portuguese. The format paper should be A4 with single column, 3.5 cm for upper margin, 2.5 cm for bottom margin and 3.0 cm for lateral margins, without headers or footers. The main font must be Times, 12 point nominal size, with 6 points of space before each paragraph. Page numbers must be suppressed.

Full papers must respect the page limits defined by the conference. Conferences that publish just abstracts ask for **one**-page texts.

2. First Page

The first page must display the paper title, the name and address of the authors, the abstract in English and "resumo" in Portuguese ("resumos" are required only for papers written in Portuguese). The title must be centered over the whole page, in 16 point boldface font and with 12 points of space before itself. Author names must be centered in 12 point font, bold, all of them disposed in the same line, separated by commas and with 12 points of space after the title. Addresses must be centered in 12 point font, also with 12 points of space after the authors' names. E-mail addresses should be written using font Courier New, 10 point nominal size, with 6 points of space before and 6 points of space after.

The abstract and "resumo" (if is the case) must be in 12 point Times font, indented 0.8cm on both sides. The word **Abstract** and **Resumo**, should be written in boldface and must precede the text.

3. CD-ROMs and Printed Proceedings

In some conferences, the papers are published on CD-ROM while only the abstract is published in the printed Proceedings. In this case, authors are invited to prepare two final versions of the paper. One, complete, to be published on the CD and the other, containing only the first page, with abstract and "resumo" (for papers in Portuguese).

4. Sections and Paragraphs

Section titles must be in boldface, 13pt, flush left. There should be an extra 12 pt of space before each title. Section numbering is optional. The first paragraph of each section should not be indented, while the first lines of subsequent paragraphs should be indented by 1.27 cm.

4.1. Subsections

The subsection titles must be in boldface, 12pt, flush left.

5. Figures and Captions

Figure and table captions should be centered if less than one line (Figure 1), otherwise justified and indented by 0.8cm on both margins, as shown in Figure 2. The caption font must be Helvetica, 10 point, boldface, with 6 points of space before and after each caption.

Figura 1. A typical figure

In tables, try to avoid the use of colored or shaded backgrounds, and avoid thick, doubled, or unnecessary framing lines. When reporting empirical data, do not use more decimal digits than warranted by their precision and reproducibility. Table caption must be placed before the table (see Table 1) and the font used must also be Helvetica, 10 point, boldface, with 6 points of space before and after each caption.

6. Images

All images and illustrations should be in black-and-white, or gray tones, excepting for the papers that will be electronically available (on CD-ROMs, internet, etc.). The image

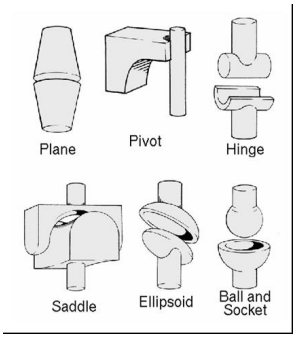


Figura 2. This figure is an example of a figure caption taking more than one line and justified considering margins mentioned in Section 5.

Tabela 1. Variables to be considered on the evaluation of interaction techniques

	Value 1	Value 2
Case 1	1.0 ± 0.1	$1.75\times10^{-5} \pm 5\times10^{-7}$
Case 2	0.003(1)	100.0

resolution on paper should be about 600 dpi for black-and-white images, and 150-300 dpi for grayscale images. Do not include images with excessive resolution, as they may take hours to print, without any visible difference in the result.

7. References

Bibliographic references must be unambiguous and uniform. We recommend giving the author names references in brackets, e.g. [Knuth 1984], [Boulic and Renault 1991], and [Smith and Jones 1999].

The references must be listed using 12 point font size, with 6 points of space before each reference. The first line of each reference should not be indented, while the subsequent should be indented by 0.5 cm.

Referências

Boulic, R. and Renault, O. (1991). 3d hierarchies for animation. In Magnenat-Thalmann, N. and Thalmann, D., editors, *New Trends in Animation and Visualization*. John Wiley & Sons ltd.

Knuth, D. E. (1984). *The TEX Book*. Addison-Wesley, 15th edition.

Smith, A. and Jones, B. (1999). On the complexity of computing. In Smith-Jones, A. B., editor, *Advances in Computer Science*, pages 555–566. Publishing Press.

Chapter 4

Conclusions

4.1 Overall Conclusions

Resultados em reconhecimento de fala, em muitas vezes, são exploratórios. As línguas são sistemas dinâmicos, variando dado o espaço, o tempo, os grupos sociais, as situações comunicativas e o próprio falante. Tendo em vista que cada língua possui aspectos particulares, não se pode assegurar, no PLN, que métodos já testados para outros idiomas, sejam diretamente transplantados para o PB e tenham funcionamento e desempenho similar. Esta dissertação busca propor um método de elaboração de um sistema de reconhecimento de pronúncia para aprendizes de inglês, falantes nativos do PB. Um sistema desse tipo ainda não foi elaborado para o PB e os resultados, portanto, são tentativos. Como se discutiu, tal sistema é de utilidade, tendo em vista a baixa proficiência em inglês dos brasileiros, demonstrada recentemente nos índices da GlobalEnglish (2012) e da Education First (2013).

A literatura pertinente da área foi revisada e métodos que se mostraram promissores foram selecionados para integrar o projeto. Pretende-se elaborador um reconhecedor de pronúncia que seja capaz de tratar nove erros de pronúncia, provendo feedback ao usuário sobre a qualidade de sua pronúncia. Os erros foram selecionados com base nos trabalhos de Zimmer (2004), Godoy (2005), Zimmer et al. (2009) e Cristófaro-Silva (2012), assumindo-se, como pronúncia padrão, o General American (GA). A abordagem de interlíngua foi selecionada para a arquitetura do reconhecedor. O modelo acústico é, assim, alimentado com dados de fala tanto de nativos, quanto de não-nativos, aprendizes de inglês. No caso, para os dados de nativos, será utilizado o TIMIT; e para os de não-nativos, o COBAI e um corpus de leitura de sentenças foneticamente balanceadas, ainda a ser compilado. As sentenças serão extraídas de um corpus de aprendizes, o COMAprend. Um script em Python será utilizado para realizar a conversão grafema-fone das sentenças, tendo por base a pronúncia canônica das palavras registrada no CMU Pronouncing Dictionary. Na abordagem interlingual, também, o modelo

de pronúncia deve ser alimentado com as variantes de pronúncia do aprendiz, de modo a compor os chamados dicionários multipronúncia. Para isso, pretende-se utilizar o CMU Pronouncing Dictionary como base e adicionar as hipóteses de pronúncia dos aprendizes por meio de regras transformacionais. O modelo de língua será constituído por trigramas e gerado a partir da Simple English Wikipedia., de modo a apresentar um sintaxe próxima à produção do aprendiz.

Um protótipo foi elaborado de modo a avaliar a viabilidade do método ora proposto. O cronograma de execução do projeto está dentro do prazo, e as bibliotecas e os softwares necessários para sua execução (HTK, Julius, Adintool, Audacity, Praat, SoX) já foram testados na elaboração do protótipo. Uma porção do COBAI (~3h40min) foi segmentada, alinhada e analisada e utilizada para estimar o modelo acústico. Tendo em vista o grande de número de arquivos do COBAI que teve de ser desconsiderado (apenas ~1h30min do que foi segmentado pode ser utilizado), optou-se por se realizar a coleta de um corpus de leitura de sentenças foneticamente balanceadas especificamente para o desenvolvimento do projeto. De modo a simular, no protótipo, os dados que se obterão com esse corpus, um corpus de erros induzidos (~2h20min) foi gravado, segmentado, transcrito e analisado. Tal corpus também foi utilizado na estimação do modelo acústico do protótipo, juntamente com os dados do COBAI. O método de coleta e anotação do corpus real, de leitura de sentenças foneticamente balanceadas, será definido em visita técnica à Universidade de Coimbra, sob supervisão da Profa. Sara Candeias, no período de 28 de janeiro a 28 de fevereiro de 2014. O dicionário-base do modelo de pronúncia, o CMU Pronouncing Dictionary, já foi testado no protótipo, tendo sido adicionadas variantes de pronúncia para um dos erros selecionados: a simplificação silábica. Foi observado que houve um grande aumento no número de entradas no modelo de pronúncia com a adição das variantes de pronúncia: o dicionário cresceu de 1.855 palavras para 7.597. Sendo assim, é possível, ao se coligir todas as regras para os nove tipos de erros, que o dicionário cresça fortemente, tornando o reconhecimento confuso, bem como consumindo tempo e recursos computacionais. Caso isso ocorra, uma solução possível seria criar dicionários específicos para cada tipo de erro, ou solicitar a um especialista que cerceie o dicionário, eliminando as variantes que ocorrem com menor frequência. Os resultados iniciais obtidos com o protótipo no reconhecimento mostraram-se promissores.

O conteúdo do projeto tem sido publicado na web[31] de modo a dar-lhe visibilidade e angariar possíveis colaboradores. Pretende-se, também, ao final do projeto, disponibilizar um pequeno sistema de treino de pronúncia, como a prova de conceito para uso do reconhecedor de pronúncia. Uma interface vem sendo desenvolvida para disponibilizar o protótipo na web, as Figura 19 e Figura 20 trazem algumas telas de exemplo dessa interface.

[pic]

Figura 19: Interface do protótipo na web - visão geral do site e tela de captura do áudio com espectro de frequência.

| [i] | [ii] | |[pic] |[pic] |

Figura 20: Interface do protótipo na web - [i] palavra reconhecida com transcrição em formato IPA e em alfabeto adaptado; [ii] tela com texto de feedback sobre a pronúncia do aprendiz, após ele reiterar no erro.

Há também a possibilidade de se realizar visita técnica ao Laboratório de Processamento de Sinais (LaPS) da Universidade Federal do Pará (UFPA), na época de avaliação dos resultados finais da dissertação, sob supervisão do Prof. Aldebaro Klautau, co-orientador da pesquisa.

4.2 Limitations

4.3 Further Work

References

[1] Benzeghiba, M., Mori, R. D., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10-11):763–786. Intrinsic Speech Variations.

[2] Bisol, L. (2005). *Introdução a estudos de fonologia do português brasileiro*. Edipucrs.

[3] Brown Jr, W., Morris, R. J., Hollien, H., and Howell, E. (1991). Speaking fundamental frequency characteristics as a function of age and professional singing. *Journal of Voice*, 5(4):310–315.

[4] Cagliari, L. C., Laplantine, F., Editora, M. F., Brait, B., Lévy, P., Mattos, R. V., Bosi, A., Hall, E. T., da Graça Nicoletti, M., and Elias, V. M. (2002). Análise fonológica. *São Paulo*.

[5] Câmara, J. M. (1970). *Estrutura da língua portuguesa*. Editôra Vozes.

[6] Collischonn, G. (2004). Epêntese vocálica e restrições de acento no português do sul do brasil. *Signum: Estudos da Linguagem*, 7(1):61–78.

[7] Cristófaro Silva, T. et al. (2012). Revisitando a palatalização no português brasileiro. *Revista de Estudos da Linguagem*, pages 59–89.

[8] Crystal, D. (2011). *Dictionary of linguistics and phonetics*, volume 30. John Wiley & Sons.

[9] Davenport, M., Davenport, M., and Hannahs, S. (2010). *Introducing phonetics and phonology*. Routledge.

[10] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366.

[11] de Medeiros, B. R. (2007). Vogais nasais do português brasileiro: reflexões preliminares de uma revisita. *Revista Letras*, 72.

[12] EducationFirst (2011). *EF English Proficiency Index 2011*. Education First Ltd.

[13] EducationFirst (2012). *EF English Proficiency Index 2012*. Education First Ltd.

[14] EducationFirst (2013). *EF English Proficiency Index 2013*. Education First Ltd.

[15] EducationFirst (2014). *EF English Proficiency Index 2014*. Education First Ltd.

[16] Fitch, W. T. and Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 106(3):1511–1522.

[17] Furui, S. (2001). *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, New York, USA, 2ª edition.

[18] GlobalEnglish (2013). *The 2013 Business English Index & Globalization of English Report*. Pearson Always Learning, Pearson.

[19] Gordon, R. G. and Grimes, B. F. (2005). *Ethnologue: Languages of the world*, volume 15. SIL International Dallas, TX.

[20] Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752.

[21] Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

[22] Ladefoged, P. (1995). *Elements of acoustic phonetics*. University of Chicago Press, Chicago.

[23] Lakoff, R. (1973). Language and Woman's Place. *Language in Society*, 2(1).

[24] Lenneberg, E. (1967). *Biological foundations of language*. John Wiley and Sons, Biological foundations of language.

[25] McLoughlin, I. (2009). *Applied Speech and Audio Processing – With Matlab Examples*. Cambridge University Press, Cambridge.

[26] Mendonça, G. and Aluísio, S. (2014). Using a hybrid approach to build a pronunciation dictionary for Brazilian Portuguese. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association INTERSPEECH 2014*, Singapore.

[27] Mendonça, G., Avanço, L., Duran, M., Fonseca, E., Volpe-Nunes, M., and Aluísio, S. (2015). Evaluating phonetic spellers for user-generated content in brazilian portuguese. *Proceedings of IJCAI 2015 – International Joint Conference on Artificial Intelligence*.

[28] Mendonça, G., Candeias, S., Perdigão, F., Shulby, C., Toniazzo, R., Klautau, A., and Aluísio, S. (2014). A method for the extraction of phonetically-rich triphone sentences. *Proceedings of ITS 2014 – International Telecommunications Symposium*.

[29] Neves, M. H. M. (1999). Gramática do português falado. vol. vii: Novos estudos. são paulo.

[30] Robjohns, H. (2010). A brief history of microphones. http://microphone-data.com/ /media/filestore/articles/History-10.pdf. Last retrieved 11-21-2014.

[31] Rocca, P. D. A. (2003). Bilingualism and speech: evidences from a study on vot of english and portuguese voiceless plosives. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 19(2):303–328.

[32] Shrawankar, U. and Thakare, V. M. (2010). Techniques for feature extraction in speech recognition system : A comparative study. *International Journal Of Computer Applications In Engineering, Technology and Sciences (IJCAETS)*, pages 412–418.

[33] Silva, T. C. (2005). *Fonética e fonologia do português: roteiro de estudos e guia de exercícios*. Contexto.

[34] Skandera, P. and Burleigh, P. (2005). A manual of english phonetics and phonology. *Tübingen: Gunter*.

[35] Steriade, D. (2000). Paradigm uniformity and the phonetics-phonology boundary. *Papers in laboratory phonology V: Acquisition and the lexicon*, 3:13–334.

[36] Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.

[37] Umesh, S., Cohen, L., and Nelson, D. (1999). Fitting the mel scale. *Proc. ICASSP 1999*, pages 217–220.

[38] United-Nations (2014). *2014 Human Development Report – Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience*. United Nations Development Programme (UNDP).

[39] Wells, J. (2000). Overcoming phonetic interference. *English Phonetics, Journal of the English Phonetic Society of Japan*, 3:9–21.

[40] Wiese, R. (2001). The phonology of/r. *Distinctive feature theory*, 2:335.

Appendix A

Phonetically-Rich Sentences - Sample Extraction

Triphone Types: 2307 | **Tokens:** 2959 | **Type/Token Ratio:** 0.78.

1. A ilha fica tão próxima da praia que, quando a maré baixa, pode ser atingida a pé.
2. Diadorim é Reinaldo filho do grande chefe Joca Ramiro traído por Hermógenes.
3. A Sicília tem alguns moinhos ainda em bom estado de conservação que lhe dão beleza e encanto.
4. Em geral, chegaram ao Brasil como escravos vindos de Angola, Congo, e Moçambique.
5. A sardinha é um peixe comum nas águas do mar Mediterrâneo.
6. Possuem esse nome pois costumam viver na plumagem dos pombos urbanos.
7. É brilhante, doce e muito harmônico, sem presença de metal na voz.
8. Para fechar Alessandro Del Piero fez outro aos 121'.
9. Roman Polanski dirige Chinatown com Jack Nicholson.
10. A atriz sabe falar fluentemente espanhol.
11. Eles achavam Getúlio Vargas um problema.
12. Oppenheimer captura cavalo com peão.
13. Um bago tem tamanho médio não uniforme.

14. Segundo relatório da força aérea belga há confrontos com a União Soviética.
15. É irmão do também antropólogo Gilberto Velho.
16. Ganhou sete Oscar e oito Emmy.
17. Qual é minha perspectiva agora?
18. Ela é um fantasma verde, feminino!
19. Justin em seguida volta no tempo.
20. Nós fizemos um álbum do Korn.
21. Desde então Edílson é fã dessas bandas.
22. Há um só senhor uma só fé um só batismo.
23. Ivan Lins faria um show em Mossoró à noite.
24. Cresceram maior que um gato.
25. Há locações disponíveis em Tóquio no Japão.
26. Preso a um tronco nenhum lugar é seguro!
27. Hoje é professor emérito da UFBA.
28. Veio até aqui e não vai mergulhar?
29. Luís Jerônimo é um jovem rico.
30. Na hora pensei: "tenho que fazer isso?"
31. A campanha teve coordenação de Sanches.
32. A mulher que você me deu, fugiu.
33. Eu nunca tive um encontro com Bianca.
34. Homer jura vingança a Burns.
35. Beijo, me liga e amanhã sei lá!
36. Um colégio é como um ser vivo.

37. Sophie é filha de um amigo gay de Alan Greg.
38. Xuxa guarda rancor e é ambiciosa.
39. No mesmo ano conhece Aldir Blanc em Viena.
40. É um imenso painel reunindo um elenco famoso.
41. A Sé integra três belos órgãos.
42. Em ambos, Shannon conquistou medalha.
43. A terra é abundante em recursos como vinagre e óleo vegetal.
44. Faça sua escolha e bom jogo!
45. Quem é que poderia sonhar com algo assim?
46. Ela é ruiva com olhos azuis.
47. Deu a louca na chapeuzinho!
48. De onde venho e para onde vou?
49. Eu choro e sofro tormentas!
50. Um falcão pousa em um pedregulho.
51. Ninguém tenha medo, nem fraqueza!
52. É membro do grupo Monty Python.
53. A sondagem de Senna pela Benetton e a chegada à kart.
54. Isto é um negócio e a única coisa que importa é ganhar
55. Robert é um forte glutão da equipe.
56. Um bárbaro no exército romano?
57. Infância e juventude em Linz.
58. Já ir à argentina era muito bom!
59. Fiquei com inveja dele.

60. Há dragões ao redor do mundo!
61. Edmond é pai do biólogo Jean.
62. A mãe lhe telefonava às vezes.
63. Tonho é tímido, humilde e sincero.
64. André Jung ocupa um lugar central no fórum.
65. Lois pergunta: "você é um homem ou um alienígena?"
66. Sua voz é um assobio fino e longo.
67. Por isso é sempre bom conferir!
68. Celso Lafer recuperou a jóia e devolveu-lhe.
69. É próxima ao Rio Parnaíba.
70. Lendo aquilo fica bem difícil.
71. A faculdade de John Oxford até hoje possui fãs fiéis.
72. Existe uma crença moderna no dragão chinês.
73. Sean Connery já sugeriu que Gibson fosse James Bond.
74. A raiz dos dentes é longa.
75. Essa noite produziu um feito singular.
76. Fim da segunda guerra mundial.
77. –No Zorra, eu fazia humor rasgado.
78. Charles vê um homem ser morto em um tiroteio.
79. Tinham um novo senhor agora.
80. É comum ocorrerem fenômenos ópticos com estas nuvens.
81. Era um cão de pelo escuro e olhos negros.
82. Há títulos na região tcheca da Tchecoslováquia.

83. Raquel Torres vai investigar a área.
84. Clay foge e leva a jovem Jane como refém.
85. Djavan jogou futebol e hóquei no gelo na infância.
86. A origem do fagote é bastante remota.
87. Um jedi nunca usa a força para lucro ou ganho pessoal.
88. Chamavam José Alencar de Zezé.
89. Um código fonte é um sistema complexo.
90. A igreja tem um altar barroco.
91. Luís Eduardo pronunciou a senha: "esgoto".
92. Quanto ao sexo: macho ou fêmea?
93. A rádio Caxias cumpriu esse papel.
94. Roger Lion é um campeão orgulhoso que ama boxe.
95. Um outeiro é menor que um morro.
96. Hitoshi Sakimoto nasceu em Yokohama.
97. Nenhum isótopo do urânio é estável.
98. Chicago é um bairro tranquilo e festivo.
99. Hong Kong continua a utilizar a lei comum inglesa.
100. Só cinco funcionam como museus.

Glossary

Application Programming Interface

Term used in computer science to refer to a list of routines, protocols and tools for building applications. Roughly speaking, an API lists all classes, methods and functions that a given package has..