FALE
FACULDADE DE LETRAS

UFMG

Em busca do contato perdido: o uso da distância de Levenshtein para identificação de contatos linguísticos remotos entre línguas amazônicas

Dalmo Buzato & Evandro L. T. P. Cunha
Faculdade de Letras
Universidade Federal de Minas Gerais
Brasil

Sumário

Guaporé-Mamoré e Complexo Cultural do Marico

- Uma das regiões de intenso multilinguismo na Amazônia (Lüpke et al., 2020)
- Multilinguismo de pequena escala
- Área linguística do Guaporé-Mamoré (Crevels e Van der Voort, 2008)
- Semelhanças culturais e formação de complexos culturais (lado boliviano e brasileiro do rio Guaporé)
  Lévi-Strauss (1948) e Maldi (1991)

Guaporé–Mamoré e Complexo Cultural do Marico

Language	Family and Subfamily	Population	Speakers
Aruá	Tupi-Mondé	94	5
Akuntsú	Tupi-Tuparí	3	4
Sakurabiat/Mekens	Tupi-Tuparí	134	16
Wayoro/Ajuru	Tupi-Tuparí	337	1
Makurap	Tupi-Tuparí	579	55
Tuparí	Tupi-Tuparí	650	400
Arikapú	Macro-Jê Jabutí	37	1
Djeoromitxí/Jabutí	Macro-Jê Jabutí	187	42
Aikanã	Isolate	400	250
Kanoé	Isolate	310	3
Kwaza	Isolate	47	27

Guaporé-Mamoré e Complexo Cultural do Marico

Bol. Mus. Para. Emílio Goeldi, sér. Antropol. 7(2), 1991

Lista comparativa de línguas

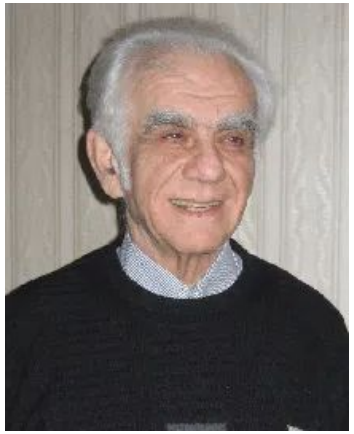
Português	Arikapú	Jabuti	Makurap	Ajuru	Koaratira/ Sakirap	Aruá	Tupari
água	bi	bzürü	ü	ügü	ükü	ü	ü
fogo	pikô	pitié	uaxát	aokap	utat	káin	kupkap
milho	titi	titi	atiti	atiti	atiti	maék	pupáp
macaxeira	boré	boré	manü	manü	tapcit	pabüiá	máin
homem	uananhé	tüê	kitô	baikop	mankup	woi	ukin
mulher	pakué	pakô	arampinhã	araminá	araminá	uazenp	araminá
civilizado	eré	eré	eré	uerep	guerep	goián	talipá
peixe	minon	minon	putkap	iboi	küpit	borip	ipot
onça	kurá	uá	amekô	amekô	amekô	nenkó	amekô
sol	tahan	tohon	gueát	jacop	tuakop	ngát	kiakop
lua	kupá	kupá	ulí	pakuri	pakuri	gatí	kuepá

Dificuldades:

- Línguas e sociedades em elevado grau de perda cultural e linguística.
- Poucos estudos descritivos e/ou diacrônicos sobre essas línguas.
- Lista de palavras em Maldi (1991) produzida sem critérios de transcrição claros e bem definidos.

A transmissão cultural acompanharia a transmissão linguística por contato?

Distância de Levenshtein

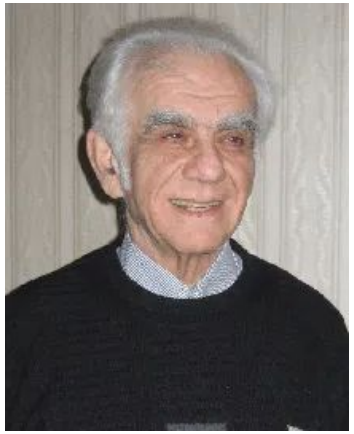


Vladimir Iossifowitsch Levenshtein (1939-2017)

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

A distância de Levenshtein é um número que nos diz o quão diferente duas strings (sequências de caracteres) são. Quanto maior o número, mais diferente essas duas strings são.

Distância de Levenshtein



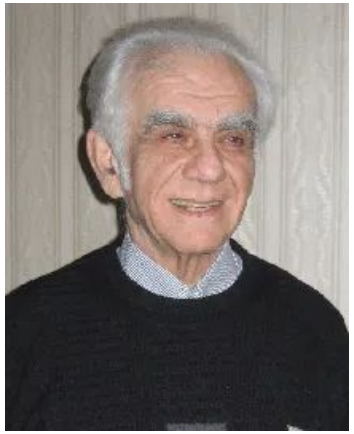
Vladimir Iossifowitsch Levenshtein (1939-2017)

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

A distância de Levenshtein é um número que nos diz o quão diferente duas strings (sequências de caracteres) são. Quanto maior o número, mais diferente essas duas strings são.

s = [‘banana’, ‘bahamas’]

Distância de Levenshtein



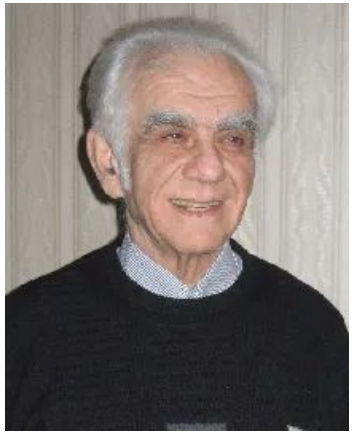
Vladimir Iossifowitsch Levenshtein (1939-2017)

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

A distância de Levenshtein é um número que nos diz o quão diferente duas strings (sequências de caracteres) são. Quanto maior o número, mais diferente essas duas strings são.

s = [‘banana’, ‘bahamas’]
dist_leven = **3**

Distância de Levenshtein



Vladimir Iossifowitsch Levenshtein (1939-2017)

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

A distância de Levenshtein é um número que nos diz o quão diferente duas strings (sequências de caracteres) são. Quanto maior o número, mais diferente essas duas strings são.

- Edição (adição, remoção, troca)
- Algoritmo muito popular em computação e NLP, criado em 1965 (verificadores ortográficos).

Distância de Levenshtein

Levenshtein Distances Fail to Identify Language Relationships Accurately

Simon J. Greenhill*
The University of Auckland

The Levenshtein distance is a simple distance metric derived from the number of edit operations needed to transform one string into another. This metric has received recent attention as a means of automatically classifying languages into genealogical subgroups. In this article I test the performance of the Levenshtein distance for classifying languages by subsampling three language subsets from a large database of Austronesian languages. Comparing the classification proposed by the Levenshtein distance to that of the comparative method shows that the Levenshtein classification is correct only 40% of the time. Standardizing the orthography increases the performance, but only to a maximum of 65% accuracy within language subgroups. The accuracy of the Levenshtein classification decreases rapidly with phylogenetic distance, failing to discriminate homology and chance similarity across distantly related languages. This poor performance suggests the need for more linguistically nuanced methods for automated language classification tasks.

Evaluating linguistic distance measures

Søren Wichmann [a,b,*], Eric W. Holman [c], Dik Bakker [d,e], Cecil H. Brown [f]

[a] Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[b] Leiden University, The Netherlands
[c] University of California, Los Angeles, United States
[d] University of Amsterdam, The Netherlands
[e] University of Lancaster, United Kingdom
[f] Northern Illinois University, United States

ARTICLE INFO

Article history:
Received 1 February 2010
Received in revised form 28 April 2010
Available online 15 May 2010

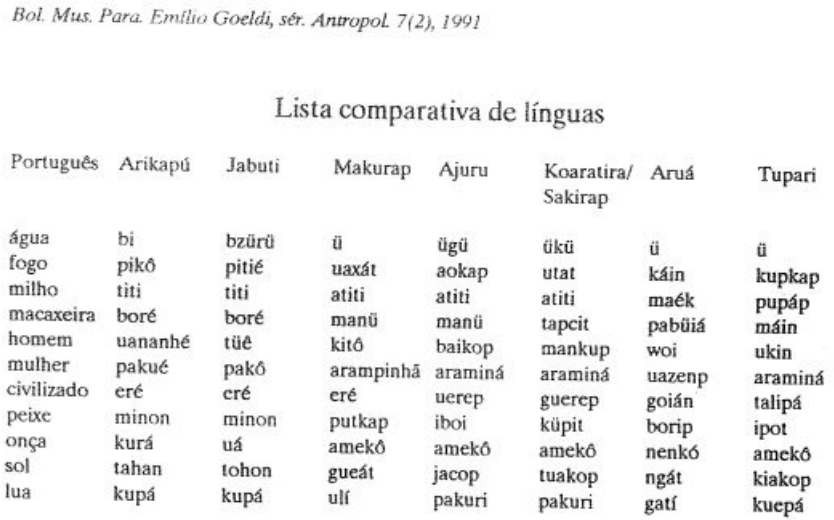
Keywords:
Historical linguistics
Phylogenetics
Levenshtein distance
Classification
ASJP

ABSTRACT

In Ref. [13], Petroni and Serva discuss the use of Levenshtein distances (LD) between words referring to the same concepts as a tool for establishing overall distances among languages which can then subsequently be used to derive phylogenies. The authors modify the raw LD by dividing the LD by the length of the longer of the two words compared, to produce what could be called LDN (normalized LD). Other scholars [7,8] have used a further modification, where they divide the LDN by the average LDN among words not referring to the same concept. This produces what could be called LDND. The authors of Ref. [13] question whether LDND is a more adequate measure of distance than LDN. Here we show empirically that LDND is the better measure in the situation where the languages compared have not already been shown, by other, more traditional methods of comparative linguistics, to be related. If automated language classification is to be used as a tool independent of traditional methods then the further modification is necessary.

E para situações de contato linguístico, como esse algoritmo performaria?

Elaboração das listas de palavras



Bol. Mus. Para. Emílio Goeldi, sér. Antropol. 7(2), 1991

Lista comparativa de línguas

Português	Arikapú	Jabuti	Makurap	Ajuru	Koaratira/ Sakirap	Aruá	Tupari
água	bi	bzürü	ü	ügü	ükü	ü	ü
fogo	pikô	pitié	uaxát	aokap	utat	káin	kupkap
milho	titi	titi	atiti	atiti	atiti	maék	pupáp
macaxeira	boré	boré	manü	manü	tapcit	pabüiá	máin
homem	uananhé	tüê	kitô	baikop	mankup	woi	ukin
mulher	pakué	pakô	arampinhã	araminá	araminá	uazenp	araminá
civilizado	eré	eré	eré	uerep	guerep	goián	talipá
peixe	minon	minon	putkap	iboi	küpit	borip	ipot
onça	kurá	uá	amekô	amekô	amekô	nenkó	amekô
sol	tahan	tohon	gueát	jacop	tuakop	ngát	kiakop
lua	kupá	kupá	ulí	pakuri	pakuri	gatí	kuepá

Lista de Maldi (1991):

- Apenas 7 das 11 línguas do complexo representadas
- Elaboração sem critérios técnicos definidos

Vantagem:

- Une elementos do vocabulário básico (água, fogo) com elementos específicos da ecologia amazônica

Elaboração das listas de palavras

LEXICO-STATISTIC DATING OF PREHISTORIC ETHNIC CONTACTS

With Special Reference to North American Indians and Eskimos

MORRIS SWADESH

Lista Swadesh:

- Lista de possíveis conceitos universais em todas as línguas para análise em glotocronologia.
- 100 itens em sua versão inicial.

Desvantagens:

- Não contém os itens lexicais da ecologia amazônica.
- Vocabulário básico tenderia a ser menos suscetível a mudanças.

Elaboração das listas de palavras

TuLaR (Tupían Language Resources)

TuLaR (**Tu**pían **La**nguage **R**esources) is an ongoing project being compiled within the CrossLingference project that collects linguistic (lexical, morphological, and syntactical) and ethnographic data related to the Tupían languages. The data is made available under CC licenses. TuLaR comprises five databases all of which are work-in-progress in different stages of completion.

Contém dados apenas das línguas Tupi

DiACL
Diachronic Atlas of Comparative Linguistics

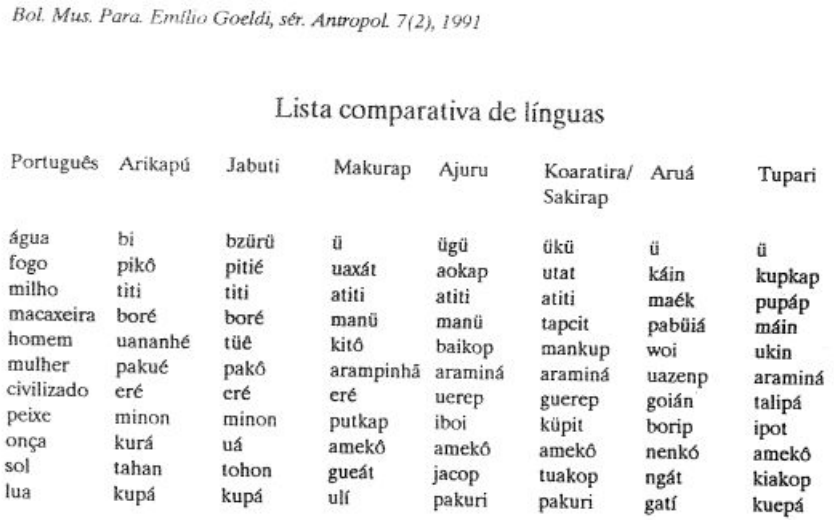
Para algumas línguas contém 120 itens, em outras, apenas 3

Welcome to The ASJP Database

The database of the Automated Similarity Judgment Program (ASJP) aims to contain 40-item word lists of all the world's languages. A lexical distance can be obtained by comparing the word lists, which is useful, for instance, for classifying a language group and for inferring its age of divergence. Click the Help link for further instructions, and for more background visit Wikipedia.

Não contém dados das línguas isoladas

Elaboração das listas de palavras



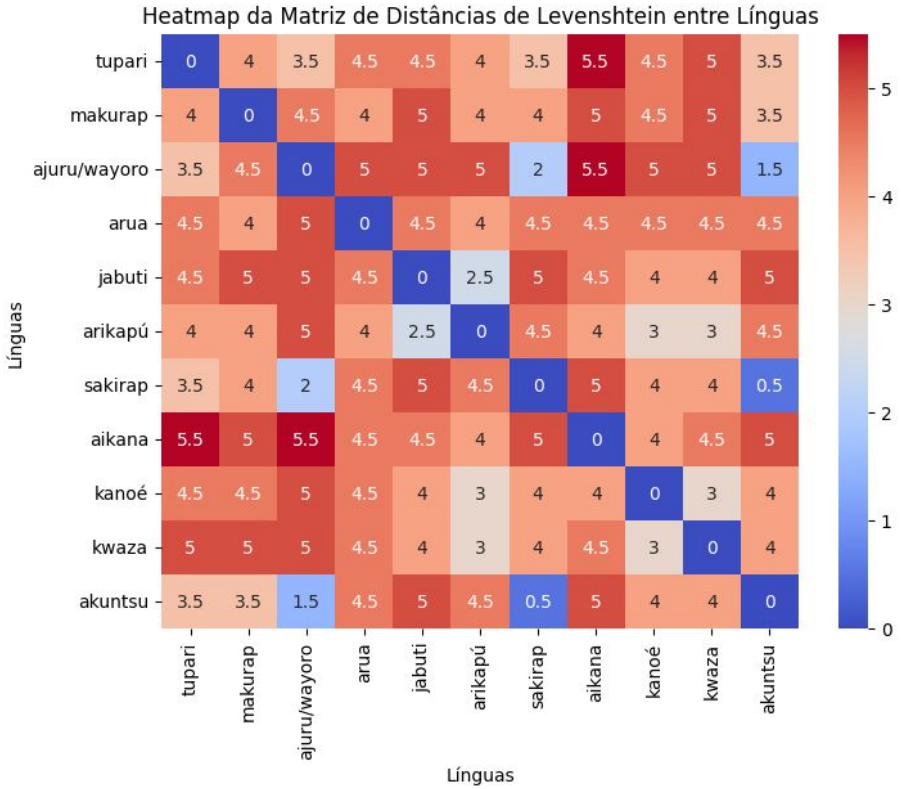
Bol. Mus. Para. Emílio Goeldi, sér. Antropol. 7(2), 1991

Lista comparativa de línguas

Português	Arikapú	Jabuti	Makurap	Ajuru	Koaratira/ Sakirap	Aruá	Tupari
água	bi	bzürü	ü	ügü	ükü	ü	ü
fogo	pikô	pitié	uaxát	aokap	utat	káin	kupkap
milho	titi	titi	atiti	atiti	atiti	maék	pupáp
macaxeira	boré	boré	manü	manü	tapcit	pabüiá	máin
homem	uananhé	tüê	kitô	baikop	mankup	woi	ukin
mulher	pakué	pakô	arampinhã	araminá	araminá	uazenp	araminá
civilizado	eré	eré	eré	uerep	guerep	goián	talipá
peixe	minon	minon	putkap	iboi	küpit	borip	ipot
onça	kurá	uá	amekô	amekô	amekô	nenkó	amekô
sol	tahan	tohon	gueát	jacop	tuakop	ngát	kiakop
lua	kupá	kupá	ulí	pakuri	pakuri	gatí	kuepá

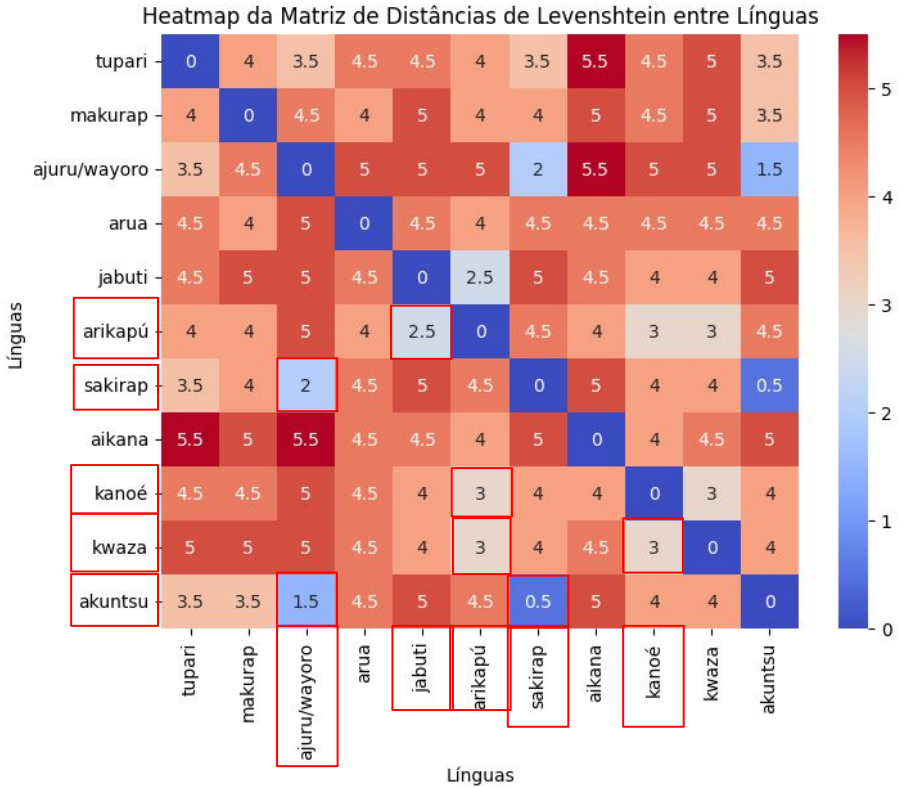
- Resolvemos utilizar as 11 palavras utilizadas por Maldi (1991)
- Para as línguas **Tupi**: TuLaR
- Para as **outras línguas (Macro-Jê e Isoladas)**: DiACL e dicionários produzidos para as línguas.

Resultados



- python-Levenshtein
- numpy
- scipy
- matplotlib
- seaborn

Resultados



- **Arikapú - Jabuti**: relação filogenética
- **Sakirap - Ajuru**: relação filogenética
- **Kanoé-Arikapú**: relação de contato
- **Kwaza-Arikapú**: relação de contato
- **Kwaza-Kanoé**: relação de contato
- **Akuntsu - Ajuru**: relação filogenética
- **Akuntsu - Sakirap**: relação filogenética

Características do multilinguismo de pequena escala e da Amazônia enquanto área linguística

- Observamos índicios de contato linguístico por meio da difusão de caracerísticas gramaticais (morfossintáticas) ao contrário de unidades lexicais (empréstimos e *code-switching*) (Epps & Salanova, 2021)

- Normalmente: 1º) difusão de itens lexicais por contato
  2º) convergência gramatical

- Este padrão não é observado em toda a Amazônia, não somente na região linguística do Guaporé-Mamoré

- Eficácia do método empregado neste caso.

Discussão e problemáticas

- Parece ser uma metodologia interessante para estudos quantatitivos e empíricos em contato linguístico, de fácil implementação computacional e análise dos dados.

  Ressalvas:
  - A qualidade e a quantidade de dados (listas de palavras)
  - Convergência dos métodos de transcrição
  - A natureza da ecologia multilíngue de contato (Croft, 2021)
  - Codificação dos caracteres

Dec	Hex	Oct	Chr	Dec	Hex	Oct	HTML	Chr	Dec	Hex	Oct	HTML	Chr	Dec	Hex	Oct	HTML	Chr
0	0	000	NULL	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	Start of Header	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	Start of Text	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	End of Text	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	End of Transmission	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	Enquiry	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	Acknowledgment	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	Bell	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	Backspace	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	Horizontal Tab	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	Line feed	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	Vertical Tab	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	Form feed	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	Carriage return	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	Shift Out	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	Shift In	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	Data Link Escape	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	Device Control 1	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	Device Control 2	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	Device Control 3	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	Device Control 4	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	Negative Ack.	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	Synchronous idle	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	End of Trans. Block	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	Cancel	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	End of Medium	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	Substitute	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	Escape	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	File Separator	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	Group Separator	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	Record Separator	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	Unit Separator	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	Del

asciicharstable.com

Tabela ASCII

#	IPA	Branner	M&O	PHONASCII	Praat	SIL	UPSID	Usenet	Worldbet	X-SAMPA	Value
157	ʎ	y&	l)	lj	\yt	L<	lj	l^	L	L	lateral approximant
155	l	l	l	l	l	l	l	l	l	l	Voiced alveolar lateral approximant
149	ɮ	l3")	l$	lz	\lz	l>	lF	z<lat>	Zl [a]	K\	Voiced alveolar lateral fricative
148	ɬ	l-	l%$	ls	\l-	l=	hlF	s<lat> or L [c]	hl [a]	K	Voiceless alveolar lateral fricative
109	k	k	k	k	k	k	k	k	k	k	Voiceless velar plosive
164	ʄ	j$	J,,	J?	\j^	j>	dj<	J`	J<	J_<	Voiced palatal implosive
108	ɟ	j-	J	J	\j-	j=	dj	J^	J	J\	Voiced palatal plosive
139	ʝ	j"	j$	J\	\jc	j<	jF	C<vcd>	j^ [a]	j\	Voiced palatal fricative
118	ɲ	nj)	n)	nj	\nj	n=	nj	n^	n~	J	Voiced palatal nasal
153	j	j	j	j	j	j	j	j	j	j	Voiced palatal approximant
											Near-close

X-SAMPA: Extended Speech Assessment Methods Phonetic Alphabet ("Alfabeto Fonético Estendido dos Métodos de Avaliação da Fala": remodelagem do AFI (Alfabeto Fonético Internacional) baseada no SAMPA que utiliza caracteres ASCII de 7 bites)

FALE
FACULDADE
DE LETRAS

UFMG

Muito obrigado!

dalmobuzato@ufmg.br
cunhae@ufmg.br

Workshop de Fonologia: Homenagem à Professora Leda Bisol

Belo Horizonte, Minas Gerais

12-13 Dezembro, 2024