

LATE fonētiski marķēts latviešu runas korpuss (fonLATE)

SATURS

Korpora vispārīgs raksturojums	2
Mašīnlasāmais fonētiskais alfabēts	2
Fonētiskā transkripcija	3
Korpora skata iestatījumi	3
Meklēšana korpusā	4

Korpusa vispārīgs raksturojums

LATE fonētiski marķēts latviešu runas korpuss (fonLATE) ir fonētiski marķētu runas datu kopa, kas iegūta no LATE-sarunas un LATE-mediji korpusa. Fonētiskais marķējums (transkripcija) ir pieejams divos līmeņos:

- 1) vārda vai segmenta standartizruna, t. i., norādīta latviešu valodas pareizrunas normām atbilstoša vārda izruna;
- 2) vārda vai segmenta faktiskā izruna, t. i., norādīta vārda izruna saistītā runā, ņemot vērā skaņu pārmaiņas vārdu sadurā, fonēmu zudumu u. tml.

Fonētiskā transkripcija pierakstīta, izmantojot latviešu valodas mašīnlasāmo fonētisko alfabētu (LVMLA). Dati no LVMLA automātiski pārveidoti starptautiskajā fonētiskajā alfabētā (IPA). Meklēt fonētiski transkribētajos datos iespējams, izmantojot gan LVMLA, gan IPA.

Pieejams: <https://korpuss.lv/id/fonLATE>

Korpusa apjoms	4 stundas (48 000 tekstvienību)
Datu publicēšanas periods	2012–2024
Izstrādes periods	2024
Izstrādātāji	Latvijas Universitātes Matemātikas un informātikas institūts
Finansējuma avots	VPP "Letonika latviskas un eiropiskas sabiedrības attīstībai" (VPP-LETONIKA-2021/1-0006)
CLARIN	http://hdl.handle.net/20.500.12574/115

Mašīnlasāmais fonētiskais alfabēts

Fonētiskajai transkripcijai izmantots latviešu valodas mašīnlasāmais fonētiskais alfabēts (LVMLA), sk. **1. pielikumu**. Mašīnlasāmais fonētiskais alfabēts tiek izmantots, lai novērstu IPA simbolu attēlošanas problēmas.

LVMLA pamatā saglabāti latviešu valodas tradicionālā fonētiskā alfabēta (LFA) simboli, t. sk. garo patskaņu un līdzskaņu apzīmēšanai.

Galvenā LVMLA īpatnība – zemsvītras izmantojums divskaņu komponentu savienošanai, piem., *piens* [pi_ens], *saule* [sa_ulex], *laiks* [la_iks], *spoks* [spu_oks], *tev* [te_u]. Savukārt garš līdzskanis tiek apzīmēts ar kolu, piem., *gamma* [gam:ax], *puse* [pus:ex].

Ir izmantoti arī vairāki SAMPA¹ (*Speech Assessment Methods Phonetic Alphabet*) simboli:

- 1) ar lielo sākumburtu E vai Ē tiek apzīmēts platais patskanis [æ] vai [æ:], piem., *zems* [zEms], *zēns* [zĒns];
- 2) x kopā ar patskaņa burtu tiek apzīmēts pārīss patskanis vairākjilbju vārdu gala zilbē, piem., *māsa* [māsax], *māsas* [māsaxs], *koki* [ku_okix], *medus* [mEduxs], *spēles* [spēlexs];
- 3) pagarināts līdzskanis tiek apzīmēts ar slīpsvītru aiz līdzskaņa burta, piem., *pasaka* [pas\akax], *mežs* [meš\];
- 4) zilbisks līdzskanis tiek apzīmēts ar pluszīmi aiz atbilstošā līdzskaņa burta, piem., *katls* [katl+s];
- 5) daļēji vokalizēts līdzskanis v vai j tiek apzīmēts ar ^ aiz daļēji vokalizētā patskaņa, piem., *vējš* [vēi^š], *govs* [gu_ou^s];
- 6) galvenais uzsvars tiek apzīmēts ar pēdiņām (") un palīguzsvars ar – procentu zīmi (%), piem., *labvakar* [%lab"vakar].

Šobrīd fonLATE korpusā izrunu iespējams aplūkot ne tikai LVMLA, bet arī starptautiskajā fonētiskajā alfabētā (IPA), piem.,

zems [zæms], *zēns* [zæ:ns];
māsa [ma:sə], *māsas* [ma:səs], *koki* [kuokī], *medus* [mædūs], *spēles* [spe:ləs];
pasaka [pas'akä], *mežs* [mef:].

Meklēt fonētiski transkribētajā tekstā iespējams, izmantojot gan LVMLA, gan IPA simbolus, sk. "Meklēšana korpusā".

¹ Mašīnlasāms fonētiskais alfabēts, kas tika izstrādāts [ESPRIT](#) projektā 1987.–1989. gadā un sākotnēji lietots dāņu, nīderlandiešu, angļu, franču, vācu, itāļu valodai, vēlāk pielāgots arī citām valodām (piem., norvēģu, zviedru, spāņu, portugāļu un grieķu valodai, 2013. gadā arī latviešu valodai), izveidojot X-SAMPA (*the Extended Speech Assessment Methods Phonetic Alphabet*). Arvien plašāk izplatoties Unicode atbalstam IPA simboliem, samazinājās nepieciešamība pēc atsevišķas, datorlasāmas sistēmas IPA attēlošanai ASCII. Tomēr X-SAMPA joprojām ir noderīgs IPA simbolu ievades ekvivalents.

Fonētiskā transkripcija

Fonētiskajā transkripcijā² atspoguļotas **pozicionālās skaņu pārmaiņas**:

- līdzskaņu asimilācija balsīguma ziņā,
 - piem., *labs* [lɔps], *aizskriet* [aizskriet],
- līdzskaņu asimilācija artikulācijas vietas ziņā,
 - piem., *mežs* [mefʃ], *aizšķērsot* [aiz:ca:rsuot],
- afrikātas [c] rašanās,
 - piem., *kāds* [ka:ts], *skats* [skats],
- līdzskaņa [j] vai [v] daļēja vai pilnīga vokalizācija,
 - piem., *zvejnieks* [zveinieks], *tēvs* [tæ:ʊs],
- līdzskaņu pagarinājums,
 - piem., *pase* [pas:ē], *pasaciņa* [pas:atstijnā],
- neuzsvērtā patskaņa kvantitatīvā redukcija,
 - piem., *brālis* [bra:līs], *brāļi* [bra:lī], *māsa* [ma:sā], *māsas* [ma:sās],
- vokāļu kontrakcija,
 - piem., *neesmu* [næ:smū],
- patskaņa [i] vai [u] daļēja vai pilnīga konsonantizēšanās,
 - piem., *maizei* [maizej], *pilijs* [piliij].

Fonēmu varianti, kas atspoguļoti fonētiskajā transkripcijā, parāda:


- pagarinātus vai garus līdzskaņus, piem., *mežs* [mefʃ], *puse* [pus:ē];
- pārisus patskaņus, piem., *žagata* [ʒagotā], *kālis* [ka:līs];
- nezilbiskus līdzskaņus, piem., *katls* [katls];
- daļēji vokalizētu līdzskaņus *j* un *v*, piem., *govs* [guoʊs], *vējš* [ve:ij];
- velāro *n* [ŋ], piem., *kungs* [kuŋks].

Korpusa fonētiskās transkripcijas līmenī “Standartizruna” (*Dictionary pronunciation*) norādīta latviešu valodas pareizrunas normām atbilstoša vārdu izruna, savukārt līmenī “Reālā izruna” ir parādīta vārdu izruna saistītā runā, ņemot vērā skaņu pārmaiņas vārdu sadūrā, fonēmu zudumu u. tml., piem.,

ortogrāfija:	<i>atbalsts</i>	<i>īstenībā</i>	<i>bija</i> [bij]	<i>ļoti</i>	<i>liels</i>
standartizruna:	adbalsc	īstenībā	bij	ļu_otix	li_els
standartizruna (IPA):	adbalsʦ	i:stɛni:bɑ:	bi̯	ʎuotʲ	liɛls
reālā izruna:	adbalsc	īstenībā	bi [^]	ļu_ot	li_els
reālā izruna (IPA):	adbalsʦ	i:stɛni:bɑ:	bi̯	ʎuot	liɛls

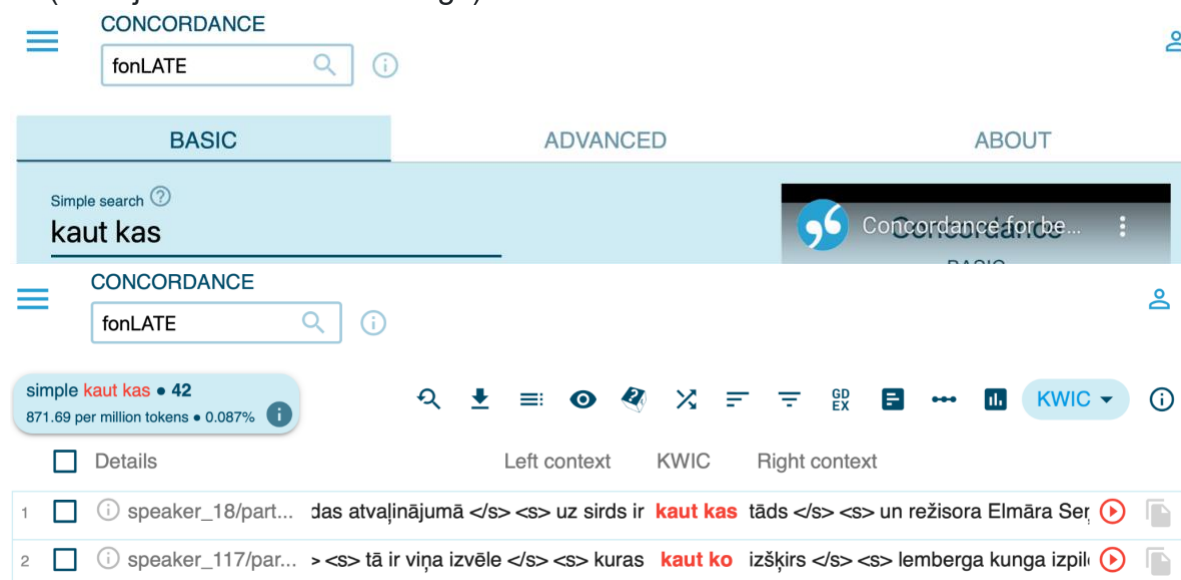
² Turpmāk tekstā piemēri norādīti IPA.


Korpusa skata iestatījumi

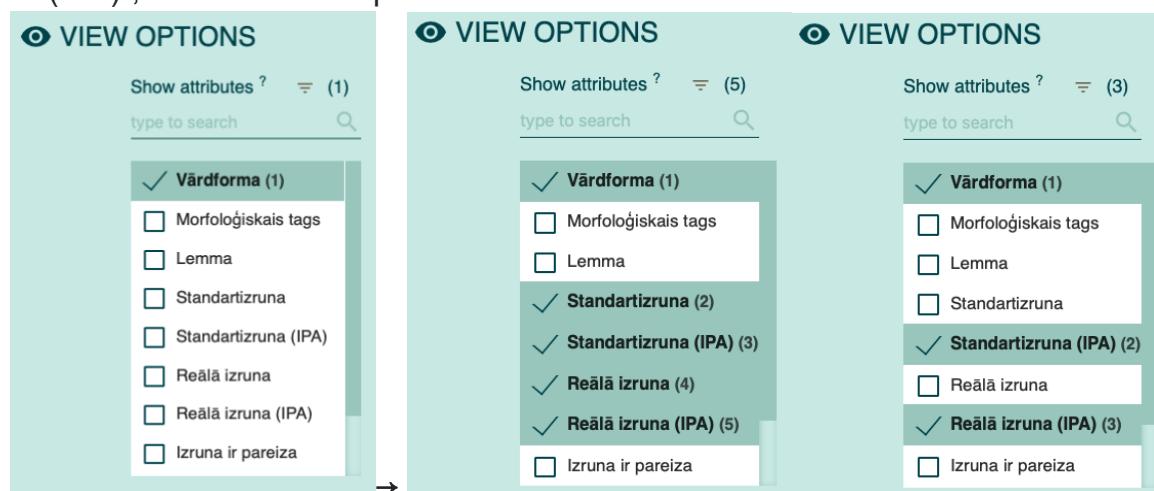
Pašlaik korpusā uzreiz nav redzama vārdu fonētiskā transkripcija. Šis skatījums ir jāiestata skata iespējās (*view options*) .

➤ Kā var redzēt fonētisko transkripciju?

1. Lai varētu redzēt abu līmeņu fonētisko transkripciju, t. i., gan standartizrunu, gan reālo izrunu, vispirms, izmantojot vienkāršu meklēšanu, ir jāiegūst konkordances (vaicājuma saturs nav svarīgs).



2. Konkordanču skatījumu var iestatīt, izvēloties  un atzīmējot iespēju “Standartizruna”, “Standartizruna (IPA)”, “Reālā izruna” un / vai “Reālā izruna (IPA)”, kā arī “Izruna ir pareiza”.



a.

b.

3. Fonētiskā transkripcija redzama zem katras vārdformas tādā secībā, kā tika atzīmēts (to norāda arī cipars iekavās), t. i., a. variantā vispirms dota

Meklēšana korpusā

Korpusā iespējams meklēt ortogrāfiskajā transkripcijā, izmantojot tieši tādas pašas meklēšanas iespējas, kādas ir, pārlūkojot jebkuru tekstu korpusu, sk. "[Vaicājumu veidi un regulārās izteiksmes noSketch Engine platformā](#)". Tā kā dati ir automātiski morfoloģiski marķēti, var meklēt, izmantojot morfoloģiskās pazīmes (sk. "[Korpuss.lv latviešu valodas morfoloģisko pazīmju kopu](#)").

➤ Kā meklēt vārdu standartizrunu?

Standartizrunu var pārlūkot, izmantojot gan mašīnlasāmo fonētisko alfabētu (LVMLA), gan starptautisko fonētisko alfabētu (IPA).

Lai meklētu interesējošo informāciju "Standartizrunas" līmenī, CQL vaicājuma logā jāraksta vaicājums **[dictionary_pronunciation="_____"]** un pēdējās jāieraksta atbilstoši mašīnlasāmā fonētiskā alfabēta simboli, piem.,

CQL: [\[dictionary_pronunciation="ka_u"\]](#)

Tiks atrasti vārda *kaut* izrunas varianti, kuros tiek atmests vārda beigu līdzskanis [t].

CQL: [\[dictionary_pronunciation=".*e_u.*"\]](#)

Tiks atrasti visi vārdi, kuros būs divskanis [eu].

CQL: [\[dictionary_pronunciation=".*t.*"\]](#)

Tiks atrasti visi vārdi, kuros līdzskanis [t] ir garš.

CQL: [\[dictionary_pronunciation=".*\(p:|t:\).*"\]](#)

Tiks atrasti visi vārdi, kuros gan līdzskanis [t], gan līdzskanis [p] ir garš.

Lai meklētu interesējošo informāciju "Standartizrunas" līmenī, CQL vaicājuma logā jāraksta vaicājums **[dictionary_pronunciation_ipa="_____"]** un pēdējās jāieraksta atbilstoši mašīnlasāmā fonētiskā alfabēta simboli, piem.:

CQL: [\[dictionary_pronunciation_ipa="kɑu"\]](#)

Tiks atrasti vārda *kaut* izrunas varianti, kuros tiek atmests vārda beigu līdzskanis [t].

CQL: [\[dictionary_pronunciation_ipa=".*eu.*"\]](#)

Tiks atrasti visi vārdi, kuros būs divskanis [eu].

CQL: [\[dictionary_pronunciation_ipa=".*t.*"\]](#)

Tiks atrasti visi vārdi, kuros līdzskanis [t] ir garš.

CQL: [\[dictionary_pronunciation_ipa=".*\(p:|t:\).*"\]](#)

Tiks atrasti visi vārdi, kuros gan līdzskanis [t], gan līdzskanis [p] ir garš.

CQL: [\[dictionary_pronunciation_ipa=".*\(p:|t:\).*"\]](#)

Tiks atrasti visi vārdi, kuros gan līdzskanis [t], gan līdzskanis [p] ir garš.

➤ Kā meklēt vārdu reālo izrunu?

Lai meklētu interesējošo informāciju "Reālās izrunas" līmenī, CQL vaicājuma logā jāraksta izteiksme **[real_pronunciation="_____"]** un pēdējās jāieraksta atbilstoši mašīnlasāmā fonētiskā alfabēta simboli.

➤ **Kā meklēt vārdu izrunu, ja mašīnlasāmā fonētiskā alfabēta simboli tiek izmantoti arī kā regulāro izteiksmju simboli?**

Vairāki regulāro izteiksmju simboli – slīpsvītra \, jumtiņš ^, pēdiņas " – tiek izmantoti arī fonēmu variantu apzīmēšanai. Lai būtu iespējams meklēt vārdu izrunu, izmantojot šos simbolus, pirms tiem jāraksta papildu slīpsvītra, piem.,

CQL: [\[dictionary_pronunciation=".*i\^.*"\]](#)

Tiks atrasti visi vārdi, kuru standartizrunā līdzskanis [j] ir daļēji vokalizējies.

CQL: [\[dictionary_pronunciation=".*t\\.*"\]](#)

Tiks atrasti visi vārdi, kuru standartizrunā līdzskanis [t] ir pagarināts, t. i., pusgarš. Slīpsvītra jāraksta gan pirms, gan aiz slīpsvītras, kas norāda uz līdzskaņa pagarinājumu.

CQL: [\[dictionary_pronunciation=".*\".*"\]](#)

Tiks atrasti visi vārdi, kuros galvenais uzsvars nav vārda pirmajā zilbē.

➤ **Kā atrast vārdus, kuru standartizruna un reālā izruna atšķiras?**

Lai atrastu vārdus, kuru standartizruna un reālā izruna atšķiras, CQL vaicājuma logā jāraksta izteiksme [\[pronunciation_correct="0"\]](#).

➤ **Kā meklēt fonLATE korpusā, izmantojot metadatus?**

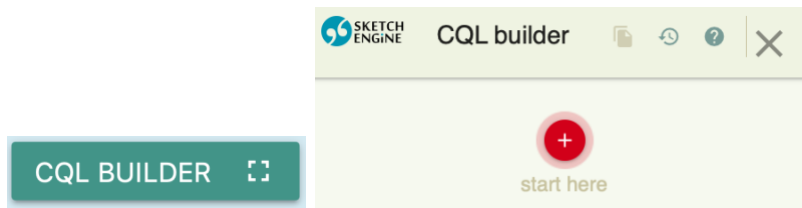
Korpusā ir iespējams meklēt noteiktā datu kopā, piem., ņemot vērā runātāju dzimumu, vecumu un latviešu valodas prasmi. Lai atlasītu interesējošu datu kopu, pirms vaicājuma ierakstīšanas, tas jānorāda, atzīmējot savu izvēli **Text types**, piem.,

Text types (2) ? ^ expand all collapse all

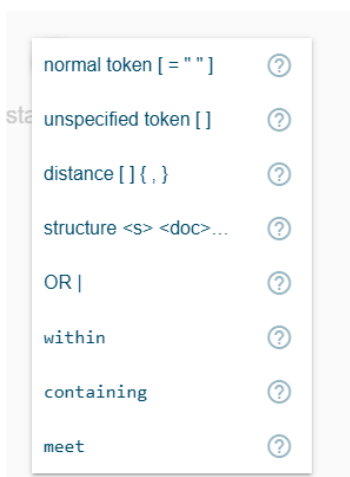
Filter	Value
Audio	▼
Dzimums	▼
Akcents	^
Dzimtā valoda (be...)	X
Bilingvāls (Divvalodīgs)	
Dzimtā valoda (ar dialekta iezīmēm)	
Nav dzimtā valoda	
Vecums	^
26-50	X
16-25	
51-65	
66-75	

➤ Kā izmantot CQL vaicājumu veidošanas palīgu **CQL BUILDER**, lai meklētu fonētiski marķētajos datos?

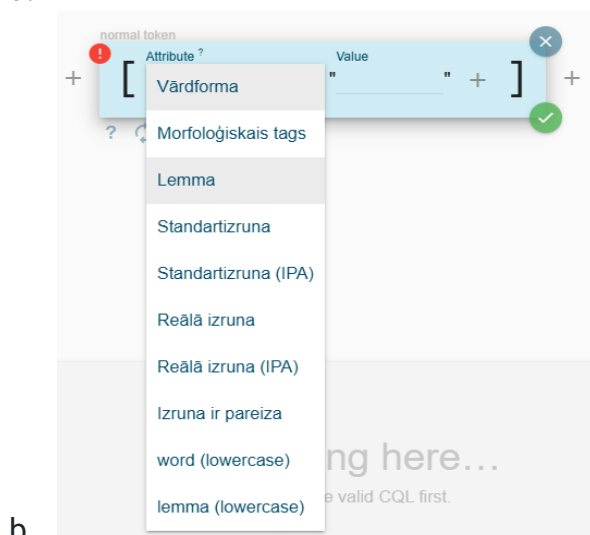
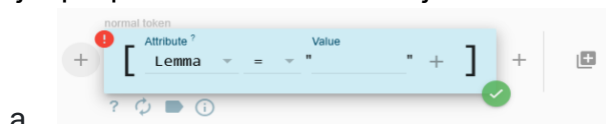
1. Jāatver CQL BUILDER, kas ir ērts palīgs / rīks lietotājiem CQL vaicājumu rakstīšanā.



2. Jāizvēlas **normal token [= " ____ "]**.



3. Atribūta laukā (a) jāizvēlas interesējošā pazīme (b), t. i., “Standartizruna”, “Standartizruna (IPA)”, “Reālā izruna”, “Reālā izruna (IPA)” vai “Izruna ir pareiza”, un jāieraksta atbilstošā vērtība, piem., [**Standartizruna = " .*ax.* "**] (c.), un jāapstiprina izveidotais vaicājums.



c.

normal token

Attribute?	Value
Standartizruna	\".*ax.* \"

? ↻ i

1. pielikums. Latviešu valodas fonētiskās transkripcijas sistēmas

Fonētiskās transkripcijas sistēmas				Piemēri			
LFA ³	IPA ⁴	SAMPA ⁵	LVMLA ⁶	Vārds ortogr.	LFA	IPA	LVMLA
Vokāļi							
i	i	i	i	<i>vilks</i>	[vilks]	[vilks]	vilks
ī	i:	ii	ī	<i>vīns</i>	[vīns]	[vi:ns]	vīns
e	e	e	e	<i>ezis</i>	[ez'is]	[ezis]	ezixs
ē	æ	E	E	<i>ezers</i>	[ežers]	[æzærs]	EzErs
ē	e:	ee	ē	<i>vējš</i>	[vējš]	[ve:iʃ]	vēi^š
ē	æ:	EE	Ē	<i>zēns</i>	[zēns]	[zæ:ns]	zĒns
a	ɑ	a	a	<i>sals</i>	[sals]	[sals]	sals
ā	ɑ:	aa	ā	<i>sāls</i>	[sāls]	[sa:ls]	sāls
u	u	u	u	<i>sula</i>	[sul ^a]	[sulā]	sulax
ū	u:	uu	ū	<i>jūgs</i>	[jūks]	[ju:ks]	jūks
o	ɔ	O	o	<i>omārs</i>	[omārs]	[ɔma:rs]	omārs
ō	ɔ:	OO	ō	<i>opera</i>	[ōper ^a]	[ɔ:perā]	ōperax
iu	i <u>u</u>	iu	i_u	<i>pliuks</i>	[pliuksķ]	[pliuksʃc]	pli_ukšķ
ie	i <u>e</u>	ie	i_e	<i>tiešs</i>	[tieš:]	[tieʃ˙]	ti_eš\
ai	a <u>i</u>	ai	a_i	<i>laiks</i>	[laiks]	[laiks]	la_iks
au	a <u>u</u>	au	a_u	<i>augšs</i>	[auks]	[auks]	a_uks
ei	e <u>i</u>	ei	e_i	<i>meita</i>	[meit ^a]	[meitā]	me_itax
uo	u <u>o</u>	uo	u_o	<i>ola</i>	[uol ^a]	[uolā]	u_olax
ui	u <u>i</u>	ui	u_i	<i>puika</i>	[puik ^a]	[puikā]	pu_ikax
eu	e <u>u</u>	eu	e_u	<i>sev</i>	[seu]	[seu]	se_u
ou	o <u>u</u>	Ou	o_u	<i>džouls</i>	[džouls]	[džouls]	d_žo_uls
oi	o <u>i</u>	Oi	o_i	<i>boikots</i>	[boikoc]	[boikots]	bo_ikoc

³ LFA – latviešu valodas tradicionālais fonētiskais alfabēts.

⁴ IPA --(International Phonetic Alphabet) -- starptautiskais fonētiskais alfabēts.

⁵ SAMPA (the Speech Assessment methods Phonetic Alphabet) – mašīnlasāms fonētiskais alfabēts, kas tika izstrādāts ESPRIT projekta laikā 1987.–1989. gadā un sākotnēji lietots dāņu, nīderlandiešu, angļu, franču, vācu, itāļu valodai, vēlāk pielāgots arī citām valodām (piem., norvēģu, zviedru, spāņu, portugāļu un grieķu valodai, 2013. gadā arī latviešu valodai).

⁶ LVMLA – latviešu valodas mašīnlasāmais fonētiskais alfabēts, kurā, cik iespējams, saglabāti LVT simboli, piem., garo patskaņu, līdzskaņu apzīmēšanai, un izmantoti atsevišķi SAMPA simboli (1) platā patskaņa e, ē, 2) pārīsa patskaņa, 3) pagarināta un gara līdzskaņa, 4) zīliskā līdzskaņa, 5) daļēji vokalizētā j un v, kā arī 6) galvenā un palīguzsvāra apzīmēšanai. Galvenā īpatnība – zemsvītra divskaņu komponentu savienošana.

Konsonanti							
b	b	b	b	<i>labi</i>	[lab ⁱ]	[labī]	labix
d	d	d	d	<i>būda</i>	[būd ^a]	[bu:dā]	būdax
f	f	f	f	<i>filma</i>	[film ^a]	[filmā]	filmax
g	g	g	g	<i>gals</i>	[gals]	[gals]	gals
ġ	ʝ	G	ġ	<i>ġērbt</i>	[ġērpt]	[je:rpt]	ġērpt
x	x	x	h	<i>halva</i>	[xalv ^a]	[xalvā]	halvax
j	j	j	j	<i>jaukt</i>	[jaukt]	[jaukt]	ja_ukt
k	k	k	k	<i>kālis</i>	[kāl ⁱ s]	[ka:līs]	kālixs
ķ	c	K	ķ	<i>ķert</i>	[ķert]	[cert]	ķert
l	l	l	l	<i>egle</i>	[egl ^e]	[eglē]	eglex
ļ	ʎ	L	ļ	<i>ļoti</i>	[ļuot ⁱ]	[ʎuotī]	ļu_otix
m	m	m	m	<i>mans</i>	[mans]	[mans]	mans
n	n	n	n	<i>nauda</i>	[naud ^a]	[naudā]	na_udax
ņ	ɲ	J	ņ	<i>ņemt</i>	[ņemt]	[ɲemt]	ņemt
Ņ	Ņ	N	N	<i>runga</i>	[ruŅg ^a]	[ruŅgā]	ruNgax
p	p	p	p	<i>puse</i>	[puss ^e]	[pus:ē]	pus:ex
r	r	r	r	<i>tur</i>	[tur]	[tur]	tur
s	s	s	s	<i>sens</i>	[sēns]	[sæns]	sEns
š	ʃ	S	š	<i>šalle</i>	[šall ^e]	[ʃal:ē]	šal:ex
t	t	t	t	<i>tur</i>	[tur]	[tur]	tur
v	v	v	v	<i>valgs</i>	[valks]	[valks]	valks
z	z	z	z	<i>ziema</i>	[ziem ^a]	[ziemā]	zi_em ax
ž	ʒ	Z	ž	<i>žagata</i>	[žagat ^a]	[ʒagatā]	žagatax
dz (3)	dʒ	dz	d_z	<i>dziesma</i>	[dziesm ^a]	[dziesmā]	d_zi_esmax
dž (ž)	dʒ / dʒ̃	dZ	d_ž	<i>džems</i>	[džems]	[dʒems]	d_žems
c	ts / t̃s	ts	c	<i>cits</i>	[cic]	[tsits]	cic
č	tʃ / t̃ʃ	tS	č	<i>četri</i>	[četr ⁱ]	[tʃetrī]	četrix
Citi simboli							
Zilbiskais līdzskanis, ortogrāfijā neatspoguļo							
◦ (ļ ļ m ņ ņ r)	◌ (ļ m ņ r)	+	+	<i>katls</i>	[katļs]	[katļs]	katl+s

Daļēji vokalizēts līdzskanis <i>j</i> vai <i>v</i> , ortogrāfijā neatspoguļo							
(<u>ū</u> i)	(<u>ū</u> i)	^	^	<i>govs</i>	[guoʊs]	[gu <u>o</u> ʊs]	gu_ou^s
Galvenais uzsvars							
._	'	“	“	<i>labdien</i>	[lab·dien]	[,lab'dien]	%lab“di_en
Palīguzsvars							
	,	%	%				
Pārīss (kvantitatīvi reducēts) patskanis							
_aeiu	ā ē ī ū	_x	_x	<i>māsa</i>	[mās ^a]	[ma:sā]	māsax
Garš līdzskanis							
līdzsk. dubultojums	:	Līdzsk. dubultojums	:	<i>lappa</i> <i>mamma</i>	[lapp ^a] [mamm ^a]	[lap:ā] [mam:ā]	lap:ax mam:ax
Pusgara fonēma				<i>mežs</i>	[meš:]	[mejː]	meš\
:	ː	\	\	<i>košs</i>	[kuoš:]	[kuoʃː]	ku_oš\