

Optical Character Recognition of Tamil Script

ஈசகீகேலளமனந்ணயவபை
ந்ணமளலககேசாயவபை
ணயவபைந்ணமளலககேசா
சகீகேலளமனந்ணயவபை

Data set 1:

Tesseract : 90.8%

Decision Tree : 72.5%

Data set 2:

Decision Tree : <20%

K. Sarveswaran

M.H.Minver

Syllables

Syllable String	Main body ID
ஈ	500
ஈ	501
ஈ	502
ஈ	503
ஈ	504

Syllable String	Main body ID
ல	505
ள	506
ய	507
ன	508
ந	509

Syllable String	Main body ID
ெ	511
ப	512
ை	513
வ	514
ய	515
ண	516

Data set 1

Main body Type	Total tokens in document images	Total unique syllables in document images
16	240	16

Data set 2

ராகுகேலளமனநனயவபெ
நனமளலககேசாயவபெண
னயவபெநனமளலககேசா
சுகிரகேலளமனநபவையண

Line Segmentation

Line:- 1

ர ச கி கே ல ள ம ன ந ண ய வ ப ரி

Line:- 2

ந ன ம ள ல க கே ச ர ய வ ப ரி ண

Line:- 3

ண ய வ ப ரி ந ன ம ள ல க கே ச ர

Line:- 4

ச கி ர கே ல ள ம ன ந ரி ப ரை வ ய ண

Ligature segmentation

ா	image_1_Li_#0_subw_#0.bmp	ச	image_1_Li_#0_subw_#1.bmp	இ	image_1_Li_#0_subw_#2.bmp	ே	image_1_Li_#0_subw_#3.bmp	க	image_1_Li_#0_subw_#4.bmp
ல	image_1_Li_#0_subw_#5.bmp	ள	image_1_Li_#0_subw_#6.bmp	ம	image_1_Li_#0_subw_#7.bmp	ன	image_1_Li_#0_subw_#8.bmp	ந	image_1_Li_#0_subw_#9.bmp
ண	image_1_Li_#0_subw_#10.bmp	ய	image_1_Li_#0_subw_#11.bmp	வ	image_1_Li_#0_subw_#12.bmp	ை	image_1_Li_#0_subw_#13.bmp	ப	image_1_Li_#0_subw_#14.bmp
ெ	image_1_Li_#0_subw_#15.bmp	ரு	image_1_Li_#1_subw_#0.bmp	ன	image_1_Li_#1_subw_#1.bmp	ம	image_1_Li_#1_subw_#2.bmp	ள	image_1_Li_#1_subw_#3.bmp
ல	image_1_Li_#1_subw_#4.bmp	க	image_1_Li_#1_subw_#5.bmp	ே	image_1_Li_#1_subw_#6.bmp	இ	image_1_Li_#1_subw_#7.bmp	ச	image_1_Li_#1_subw_#8.bmp
ா	image_1_Li_#1_subw_#9.bmp	ய	image_1_Li_#1_subw_#10.bmp	வ	image_1_Li_#1_subw_#11.bmp	ை	image_1_Li_#1_subw_#12.bmp	ப	image_1_Li_#1_subw_#13.bmp
ெ	image_1_Li_#1_subw_#14.bmp	ண	image_1_Li_#1_subw_#15.bmp	ண	image_1_Li_#2_subw_#0.bmp	ய	image_1_Li_#2_subw_#1.bmp	வ	image_1_Li_#2_subw_#2.bmp
ை	image_1_Li_#2_subw_#3.bmp	ப	image_1_Li_#2_subw_#4.bmp	ெ	image_1_Li_#2_subw_#5.bmp	ரு	image_1_Li_#2_subw_#6.bmp	ன	image_1_Li_#2_subw_#7.bmp
ம	image_1_Li_#2_subw_#8.bmp	ள	image_1_Li_#2_subw_#9.bmp	ல	image_1_Li_#2_subw_#10.bmp	க	image_1_Li_#2_subw_#11.bmp	ே	image_1_Li_#2_subw_#12.bmp
இ	image_1_Li_#2_subw_#13.bmp	ச	image_1_Li_#2_subw_#14.bmp	ா	image_1_Li_#2_subw_#15.bmp	ச	image_1_Li_#3_subw_#0.bmp	இ	image_1_Li_#3_subw_#1.bmp
ா	image_1_Li_#3_subw_#2.bmp	ே	image_1_Li_#3_subw_#3.bmp	க	image_1_Li_#3_subw_#4.bmp	ல	image_1_Li_#3_subw_#5.bmp	ள	image_1_Li_#3_subw_#6.bmp
ம	image_1_Li_#3_subw_#7.bmp	ன	image_1_Li_#3_subw_#8.bmp	ரு	image_1_Li_#3_subw_#9.bmp	ெ	image_1_Li_#3_subw_#10.bmp	ப	image_1_Li_#3_subw_#11.bmp
ை	image_1_Li_#3_subw_#12.bmp	வ	image_1_Li_#3_subw_#13.bmp	ய	image_1_Li_#3_subw_#14.bmp	ண	image_1_Li_#3_subw_#15.bmp		

Classification and recognition results for Data set 1 using DT

Type	T_Items	Correct	Accuracy
500	15	15	100
501	15	10	66
502	15	14	93
503	15	15	100
504	15	15	100
505	15	15	100
506	15	0	0
507	15	15	100
508	15	15	100
509	15	0	0
511	15	15	100
512	15	15	100
513	15	15	100
514	15	15	100
515	15	0	0
516	15	0	0
Total	240	174	72.5

Classification and recognition results for Data set 1 using Tesseract

Type	T_Items	Correct	Accuracy
500	15	15	100
501	15	15	100
502	15	15	100
503	15	15	100
504	15	15	100
505	15	15	100
506	15	15	100
507	15	15	100
508	15	15	100
509	15	8	72.2
511	15	15	100
512	15	15	100
513	15	15	100
514	15	15	100
515	15	15	100
516	15	0	0
Total	240	218	90.8

Data set 2 using line segmentation, ligature segmentation and DT

ராசுகேலளமனந்நயவபை
நனமளலககேசாயவபை
ணயவபைந்நனமளலககேசா
சுகிரகேலளமனந்நிபவையண

ே ெ ை ெ வ ை ே ெ வ ே க
ல வ ெ ை ல ல ை ெ வ ை ே
ெ ை ெ வ ல க ே வ ெ ே ை
ெ ல க ே வ ெ ே வ ை ே ெ ல
ை ெ வ ெ வ ெ ே ை வ ெ ை
ே ே க ல வ ெ ை ல

آپ کا شکریہ

धन्यवाद

ഊടിച്ചി

நன்றி