Database Specification

Overview

- Over 2200 form-pages of (946) Tunisian town/village names
- All data are digitized using a resolution of 300 dpi (b/w)
- Ground truth information available
 - o e.g. sequence of Arabic character shapes
 - o baseline/reference line position
- A wide variety of writing styles; 411 different writers
- About 26000 Arabic handwritten Tunisian town/village names
- Approximately 212 000 Arabic characters and ligatures
- Divided into 4 sets (a-d)
- Images and ground truth documentation included

Set specification

The whole database is divided into 4 disjoint sets for training and testing Arabic OCR systems. We recommend to use the specified sets for comparability with results of other groups.

SET	Number of words	Number of writer (OK+bad)
a	6537	88+14=102
b	6710	89+13=102
С	6477	88+15=103
d	6735	90+14=104
SUM	26459	355+56=411

Data formats

Image format

The form-pages are stored in uncompressed TIFF-file format (file-extension .tif).

All cropped Tunisian town/village names coming as uncompressed TIFF-images and as BMP-images (file-extension .bmp). In the TIFF header section "Image description" label information is stored.

Ground Truth format

For each cropped Tunisian town/village name you find a truth file (file-extension .tru). The truth-file is an ASCII .txt file including all available ground truth information. One example is shown below.

```
01: COM: IFN/ENIT-database truth (label) file
02: COM: http://www.ifnenit.com
03: COM: IfN, TU-BS
04: COM: di45_019.tif coming from pb377_6.tif
05: X_Y: 498 87
06: BDR: begin data record
07: LBL: ZIP:3032;AW1:مركندرويش;AW2:maB|raE|keB|zaE|daA|raA|waA|yaB|shE|;QUA:YB1;ADD:P6
08: CHA: 9
09: BLN: 56,42
10: EDR: end of data record
```

lines 01-04: comments

line 05: image size in pixel (x,y)

line 06: begin data record

line 07: label

ZIP: Tunisian post code / ZIP code

AW1: Tunisian town/village name in Arabic windows encoding

AW2: Arabic character shape sequence of Tunisian town/village name in Latin code (refer

Appendix for the lookup table).

QUA: baseline quality tag (B1=:OK;B2=:bad) ADD: number of pieces of Arabic words (PAW's)

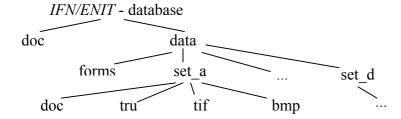
line 08: number of characters

line 09: baseline/reference line information Y1,Y2

line 10: end of data record

Database Organisation

The IFN/ENIT-database has the following directory structure.



doc

In this directory documentation in pdf and/or txt format is available.

data

In this directory all available data are stored. Each of the four sets has it's own directory (set_?).

set?

In these directories all data are available in tif and bmp file format. The ground truth information you will find in the directory tru. The subdirectory doc under data/set_? includes a pdf-file for each writer with all words and baselines.

File name convention

The following file name convention is used.

SWww NNN.EXT

- S:=set(a,b,c,d)
- Www:=writerID; W=(e,f,i,j,m,q), w=(0..9)
- NNN:=word number; N=(0..9)

Ordering Information

IFN/ENIT-database is made available for non-commercial use. The data is supplied with no guarantee of accuracy or usability. We can't guarantee to maintain the *IFN/ENIT*-database, but would be interested in hearing of any comments or results that you have.

Upon request we make the data available on the Internet for free download. If the database has to be shipped as a CD-Rom production and shipping costs will be charged. In both cases please contact us.

BUGS

When you are working with the *IFN/ENIT*-database perhaps you will discover bugs concerning the label or the extraction of the data. Please report the bugs you find back to us. So we can improve the quality of the data over the time.

Reporting results

We kindly invite you to publish reached results with the *IFN/ENIT*-database. To keep recognition results comparable, we suggest reporting results as shown in the following example:

Database version: IFN/ENIT-database v1.0p1			
Test	Training set(s)	Test set	Recognition result(*)
1	a,b,c	d	23.7%
2	c,b,d	a	25,4%

(*) Percentage of correctly recognised words in the specified test set. We recommend to use the ground truth / label category "ZIP:" as reference for the recognition result. Please use for each test the whole number of 946 different Tunisian town/village names as lexicon, in the case a lexicon is needed.

Contact

Please feel free to contact contact@ifnenit.com.

References

Mario Pechwitz, Samia Snoussi Maddouri, Volker Märgner, Noureddine Ellouze, Hamid Amiri; *IFN/ENIT-*database of handwritten Arabic words, In Proceedings of CIFED'02, Hammamet, Tunisia, 21.-23.10.2002, p.?

Appendix

Label legend / lookup table – *Arabic* label to *Latin* label & statistic of occurrence in the database

Arabic label	Latin label	quantity
0_A	0A	341
1_A	1A	280
2_A	2A	383
6_A	6A	314
7_A	7A	354
8_A	8A	285
9_A	9A	347
۶_A	hhA	520
ĩ_A	amA	544
A	aeA	1660
!_A	ahA	631
E ل_E ا	aeElaB	360
L J_B إ	ahElaB	122
M_ئ	alM	354
1_A	aaA	20256
1_E	aaE	13308

اE ل_B	aaElaB	799
M_ل E_ ا	aaElaM	1076
A_ب	baA	331
B_ب	baB	5637
<u>E</u> _ب	baE	344
<u> </u>	baM	3406
<u>ة_</u> A	teA	2184
<u>A_</u> ت	taA	358
B_ <i>ت</i>	taB	2324
ةE	teE	7261
<u></u> E	taE	357
<u> </u>	taM	1045
<u>A</u> _ث	thA	338
B_ث_B	thB	353
M_ث	thM	327
Aج	jаА	505
B	jаВ	981

Eج	jaE	346
		1219
M_ج	jaM	539
BUMج	jaMlaB	314
<u>A_ح</u>	haA	2483
<u>B_ح</u>	haB	295
<u>=</u> _ح	haE	1806
<u>M_ح</u>	haM	365
BUMح	haMlaB	64
B_ل M_م M_ح M ن B	haMmaMlaB	100
<u> </u>	haMnaB khA	341
	khB	863
	khE	321
	khM	425
	khMlaB	310
	daA	2717
ა_A ა E	daE	4886
 ა_A	dhA	353
<u> </u>	dhE	703
	raA	6255
	raE	9370
<u></u> _ A_ز	zaA	1765
<u></u> E_ز	zaE	2719
ع Aس	seA	873
—————————————————————————————————————	seB	4218
ω_E	seE	811
<u>س_</u> M	seM	1111
<u>م_</u> ش	shA	475
<u>ش_</u> B	shB	1616
<u>ش_</u> E	shE	351
<u>ش_</u> M	shM	1277
<u>A_</u> ص	saA	355
<u>_</u> B	saB	956
<u>E</u> _ص	saE	357
M_ص	saM	1096
<u>A_</u> ض	deA	351
B_ض	deB	731
<u> </u>	deE	343
M_ض	deM	328
A_ط	toA	343
B_ط	toB	359
E_ط	toE	350
M_ط	toM	1259
B_ظ	zaB	338

E_B ayB 1650 E_E ayE 34' E_M ayM 1990 E_B ghB 320 E_M ghM 600 E_M ghM 600 E_A faA 31: E_B faB 890 E_E faE 310 E_A kaA 390 E_B kaB 2600 E_B kaB 2600 E_B kaB 320 E_	M_ظ	zaM	690
E_B ayB E_E ayE S_M ayM E_B ghB S_C B ghB S_C	۸_ع	ayA	915
E_E ayE E_M ayM E_B ghB E_M ghM E_B ghB E_M ghM E_M	_		1650
E_M ayM 1996 E_B ghB 320 E_M ghM 600 E_M ghM 600 E_A faA 31.9 E_B faB 896 E_B faB 896 E_B faE 31.0 E_B faE 31.0 E_A kaA 39.0 E_A kaA 39.0 E_B kaB 2600 E_B kaE 34.0 E_B kaE 34.0 E_B kaE 34.0 E_B keE 33.0 E_B keE 34.0 E_B keE			347
È_B ghB 320 È_M ghM 600 d_A faA 319 d_B faB 890 d_B faB 890 d_E faE 310 d_A kaA 391 d_A kaA 392 d_B kaB 2600 d_B kaB 2600 d_B kaB 1300 d_B kaB 1220 d_B kaB 1434 d_B 1aB 14344 d_B 1aB 14344 d_B 1aB 14344 d_B 1aB 3889 d_B amaA 890 d_B amaA 890 d_B amaB 3889 d_B <t< th=""><th></th><th></th><th>1990</th></t<>			1990
E_M ghM 600 A faA 315 B faB 895 B faB 316 B faB 164 B A 397 B kaB 2606 B B 2606 B B 34 B B 1307 B B 122 B B 122 B B 122 B B 122 B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B <td< th=""><th></th><th></th><th>326</th></td<>			326
			600
نےE faE 310 i_M faM 164° i_A kaA 39° i_B kaB 260° i_E kaE 34° i_E kaE 34° i_E kaE 34° i_E kaE 34° i_A kaB 120° i_A kaB 122° i_A kaB 128° i_A laA 148° i_A laB 1434° i_A laB 1434° i_A laB 1434° i_A laB 1434° i_A laB 388° i_A		faA	319
ال ا	B_ف	faB	898
Jan	<u>E_</u> ف	faE	316
B kaB 2608 B kaE 348 B M 1307 B keB 1225 B keB 1225 B keE 338 B LaB 1434	<u>M_</u> ف	faM	1647
B kaB 2606 1	A_ق	kaA	397
B KaE 348		kaB	2608
ق_M kaM 130° d_B keB 122° d_E keE 33° d_M keM 98° d_M keM 98° d_A laA 148° d_A laB 1434° d_B laB 1434° d_B laB 105° d_A laB 388° d_A laB 388° d_B maB 388° d_B maB 462° d_B maMlaB 45° d_B maB 372° d_B naB 372° d_B naB 211° d_A naB 291° d_A heA 69°		kaE	348
J_E keE 33! J_M keM 98! J_A laA 148! J_B laB 1434! J_E laE 105! J_M laM 259! J_M laM 259! J_M laM 388! P_A maA 89! P_B maB 388! P_E maE 53! P_M maM 462! P_MJ_B maMlaB 45! V_A naA 126! V_B naB 372! V_B naE 211! V_M naM 291! N_A heA 69!		kaM	1307
ك_B keM 980 J_A laA 1488 J_B laB 14348 J_E laE 1050 J_M laM 2590 p_A maA 890 p_A maB 3888 p_E maE 530 p_M maM 4628 p_M maM 450 io_A naA 1260 io_B naB 3720 io_B naB 3720 io_B naB 2910 io_A heA 690	 B_ك	keB	1221
J_A laA 1489 J_B laB 14349 J_E laE 1056 J_M laM 2594 P_A maA 890 P_B maB 3889 P_E maE 536 P_M maM 4629 P_MJ_B maMlaB 456 P_A naA 1266 P_A naB 3725 P_A naB 2915 P_A heA 696	<u></u> ك_E	keE	335
J_B laB 1434! J_E laE 1056 J_M laM 2594 A maA 896 A maB 388! A maB 388! A maB 536 A maB 462! A maB 45. A maB 45. A maB 372: A maB 372: A maB 372: A maB 291: A maB 696	<u>M_</u> ك	keM	980
J_E laE 1050 J_M laM 2594 A maA 890 A maB 3889 A maB 3889 A maB 530 A maB 4629 A maM 4629 A maB 3729 A maB 3729 A maB 3729 A maB 2919 A maB 690	_L_	laA	1485
J_M laM 2594 p_A maA 896 p_B maB 3888 p_E maE 536 p_M maM 4629 p_MJ_B maMlaB 456 i_A naA 1266 i_B naB 3725 i_E naE 2119 i_M naM 2915 i_A heA 696	J_B	laB	14345
β_A maA 890 β_B maB 3881 β_E maE 530 β_M maM 4629 β_MJ_B maMlaB 450 β_MJ_B maMlaB 450 β_A naA 1260 β_B 3720 β_B 3720 β_B 2119 β_B 3720 β_B 2910 β_B 690	J_E	laE	1056
A A A A A A A A A A	<u>M_</u> ل	laM	2594
A B B B B B B B B B	A_م	maA	890
A A A A A A A A A A	B_م	maB	3885
A A A A A A A A A A	E_ع	maE	536
i_A naA 126 i_B naB 372 i_E naE 211 i_M naM 291 o_A heA 696	M_3	maM	4629
io_B naB 372 io_E naE 211 io_M naM 291 io_A heA 690	B_ل M_م	maMlaB	457
io_E naE 2119 io_M naM 2913 io_A heA 696	نA	naA	1267
<u>0_</u> M naM 291: 0_A heA 690	<u>ن_</u> B	naB	3723
o_A heA 690	<u> </u>	naE	2119
0_11 11011	نM	naM	2913
D hap 1924	ه_A	heA	696
"_B 110B	δ_B	heB	1924
	δ_E	heE	351
0_11 11011		heM	347
9_11 W 0 11	A_ و	waA	3511
<u> </u>	<u>E</u> _و	waE	6529
8_11 0.011	ىA	eeA	322
<u> </u>	ي_A		2932
	B_ي	yaB	4385
<u>G_H</u>	<u>E</u> _ى		350
	<u>E</u> _ي	yaE	2167
M yaM 7760		yaM	7760

Note:

- "IlL" is often added and means ligature "chadda"
- The character shape indicators (A,B,M,E) are sometimes supplemented with a "1" or a "2", like "baA1". In this case there is a point error detected. Take care: this feature is not consistent over the whole database ☺. We recommend ignoring these kinds of label supplements.