## **Database Specification**

#### Overview

- Over 2200 form-pages of (937) 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
  - o topline position for set\_a
- 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 using 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
c	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:مركزدرويش;AW2:maB|raE|keB|zaE|daA|raA|waA|yaB|shE|;QUA:YB1;ADD:P6
08: CHA: 9
09: BLN: 56,42
10: TLN: 23,19
11: 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

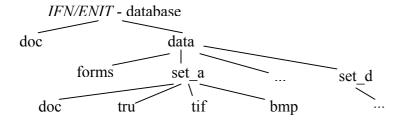
line 09: baseline/reference line information Y1,Y2

line 10: topline information Y1, Y2 (in data set a only!!!)

line 11: 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 v1.0p2				
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 937 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. 129-136

## Appendix

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

<b>Arabic</b> label	<b>Latin</b> label	Quantity
0_A	0A	342
1_A	1A	279
2_A	2A	384
6_A	6A	311
7_A	7A	354
8_A	8A	284
9_A	9A	341
۶ A	hhA	520
ĩ_A	amA	544
i_A	aeA	1660
Į A	ahA	631
ا اا	aeElaB	360
 ا E ل B	ahElaB	122
ئ Mئ	alM	355

1_A	aaA	20251
1_E	aaE	13308
ا_EJ_B	aaElaB	799
ا_E ل_M	aaElaM	1076
A_ب_	baA	331
B_ب_ B	baB	5636
E_ب_E	baE	344
<u>_</u> ب_M	baM	3407
<u>_</u> ة	teA	2182
<u> </u>	taA	356
B_ت_	taB	2324
ة_E	teE	7259
<u>ت</u> E	taE	357
<u>ت</u> M	taM	1045

ث	А	thA	338
ث	В	thB	353
<u> </u>	М	thM	327
	Α	jаА	504
<del>_</del>	В	jaВ	981
<u></u>	Ε	jaЕ	346
	M	jaM	1218
_ <u>_</u> ل M ج	В	jaMlaB	539
	A	haA	314
	В	haB	2483
	E	haE	295
_ع	М	haM	1804
_ <u>_</u> ل M _ح	<u>-</u> В	haMlaB	365
 ل M ع M ح	В	haMmaMlaB	64
ن M ح	В	haMnaB	100
<u>_</u> ع _غ	Α	khA	341
 _غ	В	khB	863
<u></u> _غ	Ε	khE	321
<u></u>	М	khM	425
 ل M خ	В	khMlaB	310
<u>و                                      </u>	А	daA	2718
ى	E	daE	4883
· · · · · · · · · · · ·	Α	dhA	353
	E	dhE	703
ر	А	raA	6252
<u> </u>	Ε	raE	9369
	Α	zaA	1764
<u> </u>	E	zaE	2718
	Α	seA	873
	В	seB	4218
<u> </u>	Ε	seE	811
<u> </u>	M	seM	1110
<u> </u>	Α	shA	475
<u></u> ش	В	shB	1617
<u> </u>	Ε	shE	351
<u> </u>	М	shM	1277
<u> </u>	A	saA	355
<u> </u>	В	saB	956
<u> </u>	E	saE	357
<u> </u>	M	saM	1096
<u>_</u> ض	A	deA	351
<u>_</u> ض	В	deB	730
 ض	E	deE	343
<u></u> ض	M	deM	328
<u> </u>		<u> </u>	1

Ь A       toA       343         Ь B       toB       359         Ь E       toE       350         Ь M       toE       350         Ь M       toE       350         Ь M       toM       1258         Ь B       zaB       339         Ь M       toM       690         € A       ayA       915         € B       ayB       1650         € E       ayE       347         € M       ayM       1990         € B       ghB       326         € M       ghM       600         € B       ghB       326         € M       ghB       326         € M       ghM       600         € B       ghB       326         € B       ghB       326         € B       ghB       388         € B       keB       <			
□       □	A ط	toA	343
b M toM 1258 b B zaB 339 b M zaM 690 c A ayA 915 c B ayB 1650 c E ayE 347 c M ayM 1990 c B ghB 326 c M ghM 600 d A faA 319 d B faB 898 d E faE 316 d M faM 1647 d A kaA 397 d B kaB 2608 d E kaE 348 d B kaB 2608 d E kaE 348 d B kaB 1307 d B keB 1221 d B kaB 14340 d B keB 1221 d B kaB 14340 d B laB 3886 d B B maB 3723 d B naB 37	 B ط	toB	359
B ZaB 339  B A A AyA 915  E B AyB 1650  E E AYE 347  E M AYM 1990  È B GhB 326  È M GhM 600  □ □ A FAA 319  □ B FAB 898  □ □ E FAE 316  □ M FAM 1647  □ A KAA 397  □ B KAB 2608  □ E KAE 348  □ D E KAE 348  □ D B KAB 1221  □ D E KAE 335  □ D M KAM 1307  □ D B KAB 1221  □ D E KAE 335  □ D M KAM 980  □ D A LAA 1485  □ D B LAB 14340  □ D E LAE 1056  □ D M LAM 2594  □ A MAA 890  □ B MAB 3886  □ E MAE 536  □ D M LAM 2626  □ M M MAM 4626  □ M M M M M M M M M M M M M M M M M M	<u></u> E ط	toE	350
B M ZaM 690 E A ayA 915 E B ayB 1650 E E ayE 347 E M ayM 1990 E B GhB 326 E M GhM 600 E A faA 319 E B faB 898 E F aE 316 E F AE 348 E F AE 347	<u></u> M ط	toM	1258
E A ayA 915 E B ayB 1650 E E ayE 347 E M ayM 1990 È B GhB 326 È M GhM 600 □ □ A faA 319 □ □ B faB 898 □ □ E faE 316 □ □ M faM 1647 □ □ A kaA 397 □ □ B kaB 2608 □ □ E kaE 348 □ □ E kaE 348 □ □ B kaB 1201 □ □ B keB 1221 □ □ E keE 335 □ □ M keM 980 □ □ A laA 1485 □ □ B laB 14340 □ □ E laE 1056 □ □ M laM 2594 P A maA 890 P A maA 890 P B maB 3886 P E maE 536 P M maM 4626 P M J B maMlaB 458 □ D B naB 3723 □ □ E naE 2119 □ □ M naM 2912 □ □ E naE 2119 □ □ M naM 2912 □ A heA 696 □ B heB 1924 □ E heE 351 □ M heM 347 □ A waA 3511 □ B waE 6529 □ B waE 6529 □ B waE 6529	<u></u> B ظ	zaB	339
E B ayB 347 E B ayB 347 E M ayM 1990 E B ghB 326 E M ghM 600 E M ghM 600 E B faB 898 E FaE 316 E M faM 1647 E M FaM 1650 E	<u></u> M ظ	zaM	690
E E ayE 347 E M ayM 1990 E B GhB 326 E M GhM 600 E A FAA 319 E F AE 316 E A KAA 379 E F AE 316 E A KAA 397 E A KAA 390 E A KAA 388 E A B MAB 3886 E A B MAB 3723 E A KAA 3724 E A KAA 3724 E A KAA 37	<u></u> A_ع	ayA	915
E M ayM 1990  E M ayM 600  E M ghM 600  A faA 319  B faB 898  E faE 316  M faM 1647  A kaA 397  B kaB 2608  E kaE 348  M kaM 1307  B keB 1221  E keE 335  M kaM 980  A laA 1485  M laB 14340  J E laE 1056  J M laM 2594  A maA 890  A maA 890  A maA 890  A maA 890  B maB 3886  A E maE 536  A maA 1267  D B naB 3723  D A naA 1267  D B naB 3723  D M naM 2912  A heA 696  B heB 1924  B heB 1924  B heB 1924  B heB 323	B_ع	ayB	
E B ghB 326 E M ghM 600  □ A faA 319 □ B faB 898 □ □ E faE 316 □ □ M faM 1647 □ □ A kaA 397 □ □ B kaB 2608 □ □ E kaE 348 □ □ M kaM 1307 □ □ B keB 1221 □ □ E keE 335 □ M keM 980 □ J A laA 1485 □ J B laB 14340 □ J E laE 1056 □ J M laM 2594 □ A maA 890 □ B maB 3886 □ E maE 536 □ M maM 4626 □ M maM 4626 □ M maM 4626 □ M maM 2912 □ D M naM 347	_ع	ayE	347
E       B       ghB       326         E       M       ghM       600         J       A       faA       319         J       B       faB       898         J       E       faE       316         J       M       faM       1647         J       A       kaA       397         J       B       kaB       2608         J       B       kaB       2608         J       E       kaE       348         J       B       kaB       2608         J       E       kaE       348         J       B       kaB       2608         J       B       kaB       2608         J       B       kaB       2608         J       B       kaB       1307         J       B       kaB       1221         J       E       keE       335         J       B       laB       14340         J       B       laB       14340         J       B       laB       14340         J       B       maA       890 <t< td=""><td>M_ع</td><td>ayM</td><td></td></t<>	M_ع	ayM	
¿ M ghM       600         A faA       319         B faB       898         E faE       316         M faM       1647         A kaA       397         B kaB       2608         E kaE       348         M kaM       1307         B keB       1221         E keE       335         M keM       980         J A laA       1485         J B laB       14340         J E laE       1056         J M laM       2594         A maA       890         A maA       890         A maA       3886         A maB       3886         A maB       3886         A maA       1267         B naB       3723         E naE       2119         M naM       2912         A heA       696         B heB       1924         E heE       351         M neM       347         A waA       3511         E waE       6529         A eeA       322		ghB	
□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □	M_غ	ghM	
□ E faE 316 □ M faM 1647 □ A kaA 397 □ B kaB 2608 □ E kaE 348 □ M kaM 1307 □ B keB 1221 □ B keE 335 □ M keM 980 □ J A laA 1485 □ J B laB 14340 □ J E laE 1056 □ J M laM 2594 □ A maA 890 □ B maB 3886 □ E maE 536 □ M maM 4626 □ M maM 4626 □ M J B maMlaB 458 □ J B naB 3723 □ D E naE 2119 □ M naM 2912 □ A heA 696 □ B heB 1924 □ A waA 3511 □ B waE 6529 □ A waA 3511	<u>A_</u> ف	faA	
J       M       faM       1647         J       A       kaA       397         J       B       kaB       2608         J       E       kaE       348         J       E       kaE       348         J       E       kaE       348         J       B       kaB       1307         J       B       kaB       1221         J       E       keE       335         J       M       keM       980         J       A       1aA       1485         J       B       1aB       14340         J       E       1aE       1056         J       M       1aM       2594         P       A       maA       890         P       B       maB       3886         P       E       maE       536         P       M       maM       4626         P       M       maM       4626         P       M       maM       4626         P       M       maM       4626         P       M       maM       1267         <	B_ <u>ف</u>	faB	
B kaB  B 1307  B NaB  B 14340  B 1	<u> </u>	faE	
B kaB 2608  B kaE 348  B kaE 348  B kaM 1307  B keB 1221  B keE 335  M keM 980  A laA 1485  A laB 14340  B laB 1436  A maA 890  B maB 3886  B E maE 536  M maM 4626  M maM 4626  M M maM 4626  M M maM 2594  B naB 3723  D B naB 3723	M_ف	faM	
© E kaE 348  © M kaM 1307  □ B keB 1221  □ E keE 335  □ M keM 980  □ J A laA 1485  □ J B laB 14340  □ J E laE 1056  □ J M laM 2594  □ A maA 890  □ A maA 890  □ B maB 3886  □ E maE 536  □ M maM 4626  □ M maM 4626  □ M maM 4626  □ M maM 2594  □ D M maM 2594  □ D M maM 2696  □ D M maM 2696  □ D M maM 2696  □ D M naA 1267  □ D M naA 12	<u>A_</u> ق	kaA	
© M kaM 1307  □ B keB 1221  □ E keE 335  □ M keM 980  □ J A laA 1485  □ J B laB 14340  □ J E laE 1056  □ J M laM 2594  □ A maA 890  □ B maB 3886  □ E maE 536  □ M maM 4626  □ M maM 4626  □ M maM 4626  □ M maM 2594  □ A naA 1267  □ B naB 3723  □ E naE 2119  □ M naM 2912  □ A heA 696  □ B heB 1924  □ E heE 351  □ M heM 347  □ A waA 3511  □ E waE 6529  □ A eeA 322	Bق	kaB	
B keB 1221  □ E keE 335  □ M keM 980  □ J A 1aA 1485  □ J B 1aB 14340  □ J E 1aE 1056  □ J M 1aM 2594  □ A maA 890  □ B maB 3886  □ E maE 536  □ M maM 4626  □ M M maM 4626  □ M D B maMlaB 458  □ D A naA 1267  □ D B naB 3723  □ D E naE 2119  □ M naM 2912  □ A heA 696  □ B heB 1924  □ A heA 696  □ B heB 351  □ M heM 347  □ A waA 3511  □ E waE 6529  □ A eeA 322	Eق	kaE	
JE keE       335         JM keM       980         JA laA       1485         JB laB       14340         JE laE       1056         JM laM       2594         AB MaA       890         AB MaB       3886         AE MaB       3886         AB MaB       4626         AB MJB       458         JA NaA       1267         JB NaB       3723         JE NaE       2119         JM NaM       2912         AA heA       696         BA HeB       1924         AB HeB       347         AB WAA       3511         AB WAA       3511         AB WAA       3511         AB WAA       3520	M_ق	kaM	
J M keM       980         J A laA       1485         J B laB       14340         J E laE       1056         J M laM       2594         A MaA       890         A B MaB       3886         A E MaE       536         A M MaM       4626         A MJ B MaMlaB       458         J A naA       1267         J B naB       3723         J E naE       2119         J M naM       2912         A heA       696         B heB       1924         E heE       351         M heM       347         A waA       3511         E WaE       6529         A eeA       322	B_ك	keB	
J A laA 1485  J B laB 14340  J E laE 1056  J M laM 2594  P A maA 890  P B maB 3886  P E maE 536  P M maM 4626  P M J B maMlaB 458  J A naA 1267  J B naB 3723  J E naE 2119  J M naM 2912  A heA 696  B heB 1924  E heE 351  M heM 347  A waA 3511  B E waE 6529  J A eeA 322	<u>_</u> ك		
J B laB 14340  J E laE 1056  J M laM 2594  A MAA 890  B MAB 3886  E MAE 536  M MAM 4626  M MAM 4626  M MAM 458  J A NAA 1267  J B NAB 3723  J E NAE 2119  M NAM 2912  A NA 696  B HAB 351  A HAA 696  B HAB 351  M HAM 347  J A WAA 3511  J E WAE 6529  J A EAA 322	<u> </u>		
J E laE 1056  J M laM 2594  P A maA 890  P B maB 3886  P E maE 536  P M maM 4626  P MJ B maMlaB 458  U A naA 1267  U B naB 3723  U E naE 2119  U M naM 2912  N A heA 696  P B heB 351  M heM 347  P A waA 3511  P E waE 6529  U A eeA 322	ل	laA	
J M laM 2594  P A maA 890  P B maB 3886  P E maE 536  P M maM 4626  P M J B maMlaB 458  U A naA 1267  U B naB 3723  U E naE 2119  U M naM 2912  N A heA 696  B heB 1924  E heE 351  M heM 347  A waA 3511  A waA 3529  U A eeA 322	<u>ل</u> B	laB	
PA maA 890 PB maB 3886 PE maE 536 PM maM 4626 PMJ B maMlaB 458  □ A naA 1267 □ B naB 3723 □ E naE 2119 □ M naM 2912 □ A heA 696 □ B heB 1924 □ B heB 351 □ M heM 347 □ A waA 3511 □ E waE 6529 □ A eeA 322	<u></u> ل	laE	
B maB 3886  B E maE 536  M maM 4626  M maM 4626  M M maM 458  D A naA 1267  D B naB 3723  D E naE 2119  M naM 2912  M naM 696  B heB 1924  B heB 351  M heM 347  A waA 3511  A waA 3529  C A eeA 322			
P E maE 536  P M maM 4626  P M M maM 4626  P M M B maMlaB 458  □ A naA 1267  □ B naB 3723  □ E naE 2119  □ M naM 2912  □ A heA 696  □ B heB 1924  □ E heE 351  □ M heM 347  □ A waA 3511  □ E waE 6529  □ A eeA 322			
P M maM 4626 P M J B maMlaB 458	\_		
PMJ B maMlaB       458         JA naA       1267         JB naB       3723         JE naE       2119         JM naM       2912         A heA       696         B heB       1924         E heE       351         MheM       347         A waA       3511         E waE       6529         A eeA       322	\ <u> </u>		
ightarranger	\ <u> </u>	_	
ightar       ightar         ightar       3723         ightar       2119         ightar       2912         ightar       696         ightar       696         ightar       1924         ightar       351         ightar       347         ightar       351         ightar       351         ightar       6529         ightar       322	· · · · · · · · · · · · · · · · · · ·	+	
ightarrow       ightarrow       2119         ightarrow       ightarrow       2912         ightarrow       ightarrow       696         ightarrow       ightarrow       1924         ightarrow       ightarrow       351         ightarrow       ightarrow       347         ightarrow       ightarrow       3511         ightarrow       ightarrow       6529         ightarrow       ightarrow       322	<u>_</u> _		
ightar       ightar         ightar       ightar <td< td=""><td></td><td></td><td></td></td<>			
o A heA       696         o B heB       1924         o E heE       351         o M heM       347         g A waA       3511         g E waE       6529         c A eeA       322			
o       B       heB       1924         o       E       heE       351         o       M       heM       347         g       A       waA       3511         g       E       waE       6529         g       A       eeA       322		+	
o E heE       351         o M heM       347         g A waA       3511         g E waE       6529         c A eeA       322			
o M heM 347  o A waA 3511  o E waE 6529  o A eeA 322	<del>-</del>		
A waA 3511  B E waE 6529  C A eeA 322			
<u>9</u> E waE 6529  Δ eeA 322	<del>-</del>		
σ A eeA 322	<u></u>		
5_11 0011			
A   yaA   2932			
	<u>_</u> ي	уаА	2932

B_ي	yaB	4383
_ ى	eeE	350
_ ي	yaE	2167
	yaM	7759

## **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.