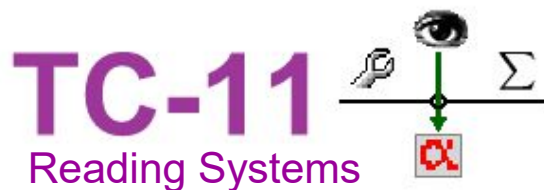




an Informal Competition in Recognizing On-line Words ICROW

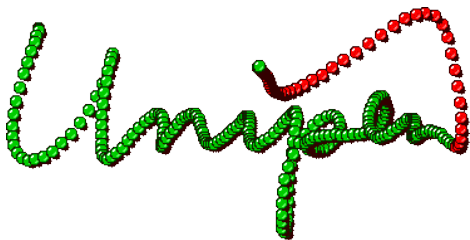
L.Schomaker



unipen.org

Overview

- Introduction
- Why an informal competition?
- The on-line data
- Images, too!
- Rules of the game

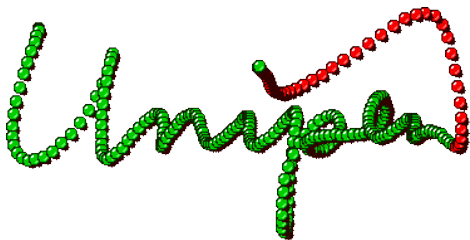


unipen.org

international
Unipen Foundation

Introduction

- International Unipen Foundation (iUF)
- Originated from IAPR/TC-11
- Idea: collect on-line data and benchmark
- Who gives, shall get in return
- Unipen data format: millions of characters by 40+ institutions

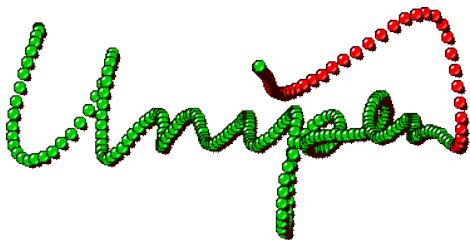


unipen.org

international
Unipen Foundation

UNIPEN Train R01/V07 set

- set up by NIST, now distributed by iUF
- Residual “unseen” data exists, too
- CDROM can be bought for a low price
- Is referred to regularly in literature



unipen.org

international
Unipen Foundation

Unipen format

.PEN_DOWN

8439 4383

2423 4390

4338 3433

.PEN_UP

4383 4389

4389 4300

4309 7874

.PEN_DOWN

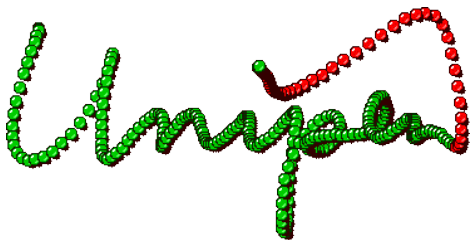
x y (z, p ...)

+ tags

writer

labels

tablet parameters etc.



unipen.org

international
Unipen Foundation

Why an informal competition?

- the UNIPEN benchmark process never materialized
- reasons:
 - cold-water fear & competition
 - difficult data
 - too many platforms/OSes

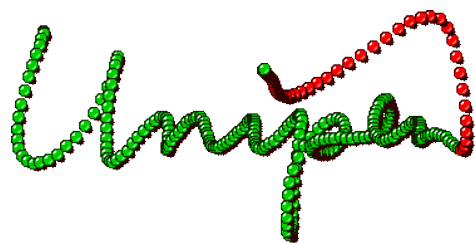


unipen.org

international
Unipen Foundation

ICROW: tests Walk-Up performance

- Data are not in the Unipen database sets
(some writer overlap does exist)
- International word list
- Variable difficulty
- Informal testing: take the data, report results. Don't train your recognizer on it.

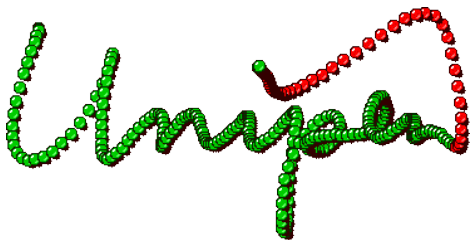


unipen.org

international
Unipen Foundation

ICROW summary:

- 884 mixed-style/case words
in dictionary: English, Italian, Dutch
- 13119 words in total
- by 72 writers, from Italy, Ireland,
& The Netherlands

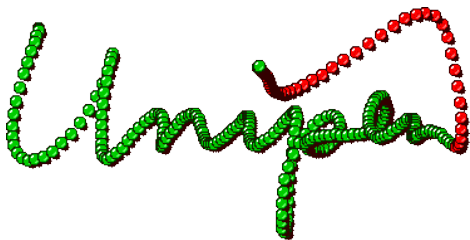


unipen.org

international
Unipen Foundation

ICROW summary:

- 31 MB unipen-ICROW-2003.tgz
- 66 MB unipen-ICROW-2003/*/*...
 - Unipen *.dat files with (X_t, Y_t)
 - Image *.png files with $I(x,y)$
- The *.png files are not annotated:
Options: TrueViz, IAM, ???



unipen.org

international
Unipen Foundation

abdomen	abstinent	adherent	adjunct	advocate	alphanumeric
aldehyde	algebra	alluvium	alp	annuensis	analyst
angst	antecedent	aorta	appendix	agua	arsen
back-p	backhinton	bangla	batik	baulhaus	bazarr
bijouerie	bladder	bolby	bodyguard	bolster	borax
bouligue	breaford	breakdown	brihano	budget	buffet
calcium	charisma	cheeklist	chevan	chloride	cochpit
colonnade	comfort	concupisio	conjunct	copywriter	cornwall
cowboy	crawl	cregut	cycle	czerny	darwin
deadline	debugger	clejuner	debi	delinquent	deodorant
disjunct	dixieland	dizzy	dozen	drunk	edukwiss
equilibrium	equipment	essay	excellent	exodus	export
exuberant	fascist	feedback	fland	fjord	flipflop
fuchsia	genre	gladiator	god	guyana	gymnast
halo	haughty	hoffman	holiday	hull	huxley
hypothesis	immigrant	inconvenient	insect	informant	inhuman
interview	israeli	island	isogues	jitter	jujube
kranchalla	keyboard	kidnapping	kuai	knockout	kravlin
larynx	lincoln	luncheon	luxo	macbeth	magister
master	maxwell	merchete	mergabels	mysterio	native
nilist	object	oh	onyx	optimum	oxford
parryus	partner	persistent	pigment	perunococcus	post
portfolio	potpourri	potdam	projector	prosperitus	quata
rembrandt	rewer	rhoso	samovar	sandwich	schizzo
showman	shuttle	sight-seeing	sleep	snob	society
squaw	stanza	stewards	stachhal	stopwatch	strychnine
stulticist	sweatshirt	symposium	tableau	teamwork	tolyo
tonc	transfer	trapezium	trekking	triplet	tuip
upstate	upgrade	vacuum	virgin	vollwater	walrus
workshop	wyoming	xylophone	yoga	yucca	zigzag

Generated by Bresenham
line generator, using
a smooth brush.

Tablet parameters are used
to scale to a size which
corresponds to 300 dpi
off-line scans.

abdomen

abstinent

aldehyde

algebra

angst

antecedent

backup

badminton

aldehyde, style "handprint"

aldehyde

Laadje

Fatsoen

Vuist

Mandje

Kien

Fourage

Figuur

Zonder

Landing

Lenig

Zender

Codeur

Veerpont

Borden

Worst

Korsten

Hortensia

Varen

Huisje

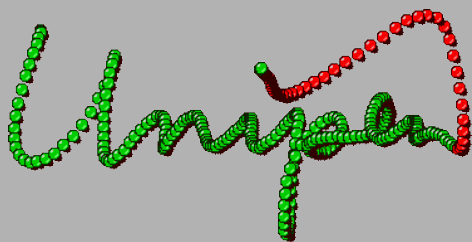
Hinder

Zien

Citrus

Banden

Hagedis



unipen.org

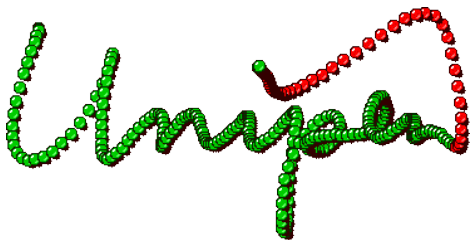
international
Unipen Foundation

Freq	Label	Lang
43	"object"	E+NL
5	"obsoleta"	I
5	"oceano"	I
9	"of"	E+NL
5	"offer"	E+NL
43	"ohm"	I+NL+E
5	"omogenei"	I
9	"on"	E
9	"one"	E
43	"onyx"	E+NL
5	"open"	E+NL
5	"opportuno"	I
43	"optimum"	E+NL
9	"or"	E
5	"orbit"	E
5	"order"	E+NL

More rules

- lexicon = all words in combined set
- report per data set

- details need to be sorted out
 - Use Simon Lucas's evaluator code?



unipen.org

international
Unipen Foundation

Address

- prof. dr. Lambert Schomaker

AI Institute

Groningen University

9712 TS Groningen

The Netherlands

schomaker@ai.rug.nl



unipen.org

international
Unipen Foundation