Recognition and information extraction in historical handwritten tables: toward understanding early 20th century Paris census

Thomas CONSTUM¹, Nicolas KEMPF¹, Thierry PAQUET¹, Pierrick TRANOUEZ¹, Clément CHATELAIN², Sandra BREE³, François MERVEILLE⁴

> 1: LITIS EA4108, University of Rouen Normandy, France 2: LITIS EA4108, INSA Rouen Normandy, France 3: LARHRA, UMR 5190, CNRS, France 4: Campus Condorcet - GED, France

> > DAS, 23rd May 2022

Olitis





Introduction

- The 20th century census of Paris (1926, 1931 and 1936) contain information on approximately 9 million individuals in total.
- Demograph historians could use the content of these census to answer questions such as:
 - What is the proportion of divorced, married, or cohabiting individuals in Paris in 1926 by district?
 - How did the structure of households in Paris evolve between 1926 and 1936 in terms of number of individuals and number of children?

	-			au uu uu										And
Pres ,		Quenor	Hinto 1				the l	Red 23	4 %	lade mark	2 41	fine	2	*
100		i lavergra Richau	The .	11 Handel 11 /4 g.	4	and Ball	1.24 C	(dut)	4 Ha	ralist our	der 22 3	9 and		
		: Geaton Gohllow	And .	4 9° time 11 fr.	- d.	- ging	ush a		4 50	the for the second	- 11	9 7		- U.X.
	11	, Dilata	and and a	1/1	- 2	5.40	NJPL -		1 40	errier to	4 11 1	nin .	4	-44 (43) Eur
		y maile	popul	11	- 7.	andie	and the	1.1	* 9	where the	14 11	0		alle 23
		1 Chiopunor		111	4	7.00	175		1 7	it in	- 11	D red	- +	A MS
		3 Machant	Horiza	1 1 The Tang	a d d	-	Cher C	Att de l'	. , 0	Bala 4	4 12) 124 14	67 67	- 4	
		1) Barbala	- ale	1 8 4	lear of		ester pe		4 4	iale de	da 12, bas 12	T.		- US
		1 carre	formation .	4 8 gu	4	- Ann	user o	1		1. 4	- 11 7 11	67	1	1 200 10
	21	" Thank	and and	11 Jaca	4	, istin	15 genet	To alla	- 3	lemen G	e 11 in 11	I and	12	
		1.		11 00 00	- 4		di di			qui tur Como de	- 17 c 44 N	margen	4	and a
		1 Edger	Heren .	11 and	· 4.		EUMO E		1 10	angini dali ampir di	11	and the second	- 4	my man \$1
7		, and	Helin Talana	31 1 A.	· +=-		22.00	-	- + 4	lanne da	3. 11 1 11	14 and		my U
ľ	22	1) Subre	12	10000		antes Aur	4		4	1	44 4	1	- 4	
		. Tunal	Hine .	11 1. 1. 11 9 dame	- 72		HISPL D		1100	2mg &		0 quy	512	lay of the

Example of a double page from the Paris census. (1926 census - Gaillon district)

Introduction

The POPP project (Project of Ocerization of the Parisian Population)

Aim: Get tabular data from double page scans in order to create a database containing information of about 9 million individuals.

AND ADDRESS AND ADDRESS ADDRES	no childra anni anni anni anni anni anni anni an	1943-1940 1.007 3.000 0 0 10 1943-1940 1.000	ALL ALL AND ALL ALL ALL ALL ALL ALL ALL ALL ALL AL	Ana in general del de la constanti del del del del del del del del del del	10001000 10001000 10001000 10001000	HARDS IN FREE FLAM Annu And	Pathaga de Lancad	um vento en el como en	ARCHARY THE ARCHARY THE ARCHARY ARCH	Anne the protection of the interaction of the interaction of the interaction of the inter- tion of the interaction of the interaction of the inter- section of the inter- mental of the inter- tion of the inte
Price 14	Quenot	Hender 11		- ale y	Prie di kontine	22 y Valade	mental sy	find -	the second	1832
(Lab)	4 Carerges	The of Gental	4 493	4 440	(Inste)	2 Rovalist	Mayyelor 22	D Testa	d again	12 sper
	1 Chapon _	This of & human	de explo	a 11393 0	17-22-54	4 Contraci	Marie IT	D the	- innit	12/103
	" Gohland	Inquel 11 ft.	n de - gin	L ann 3		4 Donale	Junger 11	14.20	d. Jan	11394
11	, Zelatte	have 1/1	- 1 2.6	# HISPIE	10 300	4 Leconorise	Smith 11	min	d alit	14362
		sens 1 1	- + +	1071 4	14.0	4 Fredor	Think 11	A have -	el seither	. 4535
	2 Reular	and it	d. ay.	60 11373 36		14 Lapobian	4 Laborator 11	O there .	a you	1 13.14 2
	, Cormank	1 1/1-	4 5	155		a marticle	Halline II	1) puterle	d. min	a uses t
	2 Willing -	John 5" (Andigen Forming 11 (France	a d d - m	- Empo	Ret de	S . + Oucher	Afred St.	100 .	- 4	× 2333
	3 148	mys Is & highly	w and by	a win o	Lepusie	1 Geals	- Marthald II Mandan II	lor !!	4 4.3	e uspe
	11 Barbarans	gile 11 & high	a di dada	4 Marcia	340		Talme 11	ar.	· * ····	11394
	1 Interroporte	much 4 & give	4. 1		-		Here 12	67	1 42	16374
	" Hold	Beaud at & human		an 11321-E		- 1	Hac of	lor	1 1	12 spli
	17 These	Bune 1 & Selimite	egen a de an	en 12340 tot	in parts	- Zamo	2 Carine II Heilana II	9 and OT 19	d pide	note U spage
		man H here to	- 4	att		- Chemi	u yaya 11	height	- d alla	64U332
	1 Edgar	Fernian 11 a.e.a.	- 4 - +	0		1) Margin	hilmite II	Cando Harrando	- 4 -	en sist.
	1.	Gongth 11 Hughlow	- 4	- ELANIA	- Caro	- 7.	Chine 91	1994		· 45397
7	. Gineslet	tedanie II hagere	d. d	an ester	- A State	9 y Jann	Bern Jo dala 6	18 rend		
	1 Secon	Jung CO Roma	nie / de - a	inter -		4 th	respla to	14000	- 4	
22	7 242400	time 11 - t.	- # +	- ering 2		×	- din to	A que	4. 1.	in hist

Example of a double page from the POPP corpus. (1926 census - Gaillon district)

		14			Contraction of the local data and				Situation par	and the second se		
			Annee de	Ville de	Departement de			Etat	rapport au chef	Statut		Code
Noms	Prenom	Sexe	naissance	naissance	naissance	Pays de naissance	Nationalite	matrimonial	de menage	profession	Profession	metier
VALADE	MARTIAL	Н	1909		GIRONDE	FRANCE	FRANCAISE	C	CH	APP	APPRENTI CUISINIER	18371
CUERVO	RAOUL	н	1893			CUBA	CUBA	C	CH		INGENIEUR	12174
KOVVALISKI	\$MIEEZYSLOV\$	A	1872			POLOGNE	POLOGNE	C	CH		NEGOCIANT	18374
CONTUCCI	MARIO	н	1889			ITALIE	ITALIE	M	CH		PROPRIETAIRE	22408
CONTUCCI	SCORRINAS	F	1887			ITAL IE	ITAL IE	M	EP	-	SANS	22408
DONATO	GEORGES	H	1896		ALPES-MARITIMES	FRANCE	FRANCAISE	C	CH	D	DOMESTIQUE COMMERCIAL	18374
I ECORVOISIER	\$CAMELLES	A	1895		NIEVRE	FRANCE	FRANCAISE	C	CH	-	ARTISTE	19388
BROVN	ROSE	F	1889			AMERIQUE	AMERIQUE	M	CH		SANS	22408
ESCHOPP	THEOPHILE	Ĥ	1872			SUISSE	SUISSE	M	CH		VEILLEUR	21395
CAPOBIANCO	EDUARDO	н	1889			ITALIE	ITALIE	M	CH		NEGOCIANT	18374
CAPOBIANCO	ANNA	F	1889			ITALIE	ITALIE	M	FP		NEGOCIANT	18374
MARTISKA	CATHERINE	F	1898			TTT Na Ta	TTT No. To.	C	CH		CUISINIERE	21300
BECKER	ALERED	H	1852	DARIS	SEINE	FRANCE	FRANCAISE	M	CH	-	CONCIERCE	21300
BECKER	MATHILDE	F	1851	17445	JEINE	THOUGH	THUNGINGE	M	EP	1	CONCIERCE	21390
CEALIS	CLAUDIUS	H	1872	DARIS	SEINE	FRANCE	FRANCAISE	M	CH	-	REPTOCE	1837/
CEALIS	DALILINE	E	1992	DARIS	SEINE	ERANCE	ERANCAISE	M	ED		SANS	1937/
CEALIS	CLAUDE	H	1902	PARIS	SEINE	FRANCE	FRANCAISE	C	F	-	REPTOCE	1837
CEALIS	GUY	H	1904	DARIS	SEINE	ERANCE	ERANCAISE	C C	E		REPTOCE	1837
CEALIS	MAX	H	1907	PARIS	SEINE	FRANCE	FRANCAISE	C	F		REPTOCE	1837
HAMON	LOUISE	E	1887	PAILO	COTES-DU-NORD	ERANCE	FRANCAISE	C	D		CLUSINIERE	2130
7515	EMILIENNE	E	1907		DAS DE CALAIS	ERANICE	ERANICAISE	C	D	-	EE CHAMPPE	2120
THOMAS	GEORGES	- L	1906		PAS-DE-CAEAIS	MADAGASCAR	MADAGASCAP	C	D	-	VALET CHAMPPE	21200
MAZZINI		E	1996		LANDES	ERANCE	ERANCAISE	C	CH	E	EMPLOYECCE	1937
SOLIVEIS	LOUIS	L	1990		LEDALIT	EDANICE	EDANICAISE	M	CH	E	EMPLOYE RUPEAU	1007
SOUVER	LOUISE	E	1902		SEINE ET MADNE	EDANCE	EDANCAISE	M	ED	L .	CANE	1037
GANNE	DIEDDE	- F	1956		COTES DIL NORD	EDANCE	EDANICAISE	M	CH		CONCIERCE	21200
GANNE	SODUIE	E	1962		SAONE ET LOIPE	EDANCE	EDANICAISE	M	ED	-	CONCIENCE	21395
LOR	ADOLDHE	F U	1002		AIGNE	EDANCE	EDANCAISE	M	CH		SANS ENEDECE	
LOR	ELOPE	-	1970	APC	ADS	EDANCE	ERANCAISE	M	ED	APC	SMEDESS	APC
DEDEZ	SELIDIOLIES	A	1009	AbS	AD3	ECDACHE	ESDACHE	m C	CH	AD3	EMDLOVE CCE	1027
PEREZ	FONDER		1070	ALCED		ALCEDIE	ALCEDIE	ŭ	CH	E	EMPLOTE CCE	1037
PRETOT	LOUICE	n	1072	ALGER		ALGERIE	ALGERIE	M	CH		NEGOCIANT	1037
PRETOT	LOUISE	F	1005	ALCER		ALCEDIE	ALCEDIE.	M	EP		SANS	1037
FEDDACCON	DOLANDE	-	1905	ALGER		ALGERIE	ALGERIE	C			SANS	2120
FERRASSON	ROLANDE	P	1909	ALGER		ALGERIE	ALGERIE	C	D		FECHAMBRE	2139
CLEMENT	ALBERTINE	F	1894	ALGER	410115	ALGERIE	ALGERIE	U.	U		CUISINIERE	2139
GOUJON	LOUIS	н	1873		AISNE	FRANCE	FRANCAISE	M	CH		COMPTABLE	1837
JOUJON	LUUIS	н	1873		AISNE	FRANCE	FRANCAISE	M	CH	-	COMPTABLE	1837
JOUJON	MARIE	F	1874		AISNE	FRANCE	FRANCAISE	M	EP	100	SANS	18375
JOUJON	SRENES	A	1908		AISNE	FRANCE	FRANCAISE	C	F	APP	APPRENTLEBENISTE	11150
ARIVE	FERNANDO	H	1900		LOID ET OUES	ESPAGNE	ESPAGNE	C	CH	E	EMPLOYECCE	1837
MARAIS	CONSTANT	H	1859		LUIK-ET-CHER	FRANCE	FRANCAISE	C III	CH		MD HOTEL	183/3
NOYAL	ROBERT	H	1894		NUC (DC	FRANCE	CD ANOLUOT	M	CH	P	PATRON FLEURISTE	9139
NOYAL	LOUISE	F	1895		NIEVRE	FRANCE	FRANCAISE	M	EP		PATRON FLEURISTE	9139
NOYAL	DENISE	F	1925			TALIF	TALIE	C	F	-		0110
MARIA	\$MIRRA\$	A	1883			TIALLE	TTALLE	M	CH	0	OUVRIER COUTURIERE	91132
ARTHEZ	RAOUL	H	1887			BRESIL	BRESIL	C	CH		SANS	22408
MELOUI	\$ERNESTA\$	A	1851			TIALIE	TIALIE	V	CH		SANS	-

Corresponding result in the POPP database.

The POPP corpus



Example of a double page from the POPP corpus. (1926 census - Gaillon district)

- 3 different census: 1926, 1931 and 1936
- For each census:
 - 100 000 pages
 - 3 million individuals
 - 80 districts
 - between 80 and 500 writers
- Handwritten tabular data with for each individual:
 - 10 columns
 - one row

The POPP processing pipeline



Diagram of the POPP processing pipeline.

Pre-processing steps: information localization



Diagram of the POPP processing pipeline.

Information localization: from double pages to tables

- Segmentation of the tables located in the image using dhSegment¹, a pixelwise predictor, which assigns the class "table" or "background" to each pixel of the image.
- Dewarping of the segmented quadrilaterals to obtain straight tables.

MID: 100: With the second	And statement between sta	
The a failed to the stand of the standard of a	ore the Cine 22 4 Jalach muld of Fine	2 min 18312 1 2 min 1842 12
7) Friderice much 11 / 19 g - 4 - 4 7) Stapper Pari 4 9 triver - 4 - 4 1) Gebelande hugut 11 / 7 4 -	altan 16 17 0 6 16 11 13 13 0 ging 11 19 18 ging 11 19 18 19 19 19 19 19 19 19 19 19 19 19 19 19 1	4 month the the second
11 y Zalatta La 1 1 1 - d	1. Ush 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
19 Mailar poul 1/1 d. 19 Auguran 1/1/1 d.	unten 1825 1 9 Societar Hinde 12 D Anna - 9 Etal 1823 1 9 Happenaus Linna 11 D Fake -	al valler 1932 to al years 1922 to to
1) Theoremands 1 1 1 1 1. 2) Millione John 67 (6) humingue al. 2) Theoland Theoreman II Theoreman al. al.	an Const C 24 de 9 4 Perler Afed 12 Tot -	a may 233
2) Helt popl 21 the highlaw and -	hangen 1232 10 " " franker T. Ballet I' OT -	- 4 - 4 - 4 - 4 - 4 - 4 - 4 - 4 - 4 - 4
1 Zalon - Zecture II nyphin de 4 portemporto Birch de Digitar de 11 houte Band 21 De houseport de	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 non using
21 19 Thannon adat 19 Isteam at 19 The same at 19 These The Ist and the adat	When 13 22 2 notice 13 33 4 the - Thermore Cause 11 & nod 	a marine 1999
1. and 1. 200	aft	d and the state
1. Grangette 10 Angellen - 4 1. Sellen 11 12 - 4an Grandelle Stamme 11 Augura de-	an What I g game Ban I ban	- 4 and 18391 -
" Factore Juge 17 D. Dominie 1 d. 22 7 Sandhard Free 11 D. Dominie - d	antes - 2 2 table 12 tables	a de mate and a second
" Turnellag Harder 2" & faither de	soin 11.521 D . I Tang Ruge of D Sugar	1 and the state of

Double page scan.

								-	and the second		1.1.1	1				1.	-	ŝ .
14		Lanot	Head	11		-	+	-	ale	d'halen	22	#	Valade	mental	4	1 gund	4	41-air
2	ų.	arrigne	These	N	Kentel		di	420	1-+	Sheek)		4	Sucros	Cand	11	@ Sula	- ú	y
	2/	Rudan	Marcid	.95	1+9		4	a palace	4.27	2		2	Romad	mayyalow	22	D Ordiges		-yes
	1	Shapon.	Ani	4	7 his		4	4 62	11.513	1		4	Sontucce	Mair	"	D Hade		report
		Gonland	high	11	pr	-	4	- gint	usp				5 .	Source	11	1 -	- +	- met
		2.4	Medan		del.	-	4		usn			4	Donale	grage .	14	14. 117	4	and and
16	y.	Lacona	Same	E		-	4	2.00	11372			1	Reconstan	Served	1	aria	4	alit
		1. J.	Maria	1	1.	-	*		asm			-	B	11 11	11	Ch.		++++
	2	Really	me		1.		4.	and the	MATAN			1	Audian	41 -	14	A 200	- 4	3-6-Chus-
		- Augunan			1.		"		in	1			1	-	"	Q d		7.
		Million	1	1	0 00	. 1	2	an	11 100			14	martiska	Fathere	41	A stal	12	
		Luberd	an	11	A E	-	2	-	22 Mor	244	0	+	Pecke	Aller	a	100		
	2	Table		2	a Li	fore L	4	here	12354	Lynam	1		1.	millit	3	0		1
	v	Fartulne	gula .	11	3 mg	he -	d.		22hor	14		y	Reals	Standard	11	100	- 4	41.92
	1	Laton	quites.	11	P into		d.	Lings	11.195	1000			1	Tankina	11	OP.		-
	9	Antonopoulo	marche	4	Q quin	-	4.	fre	-	-			- 1	Hend	12	100	+	4.0.
	-	Koli	Barnad	01	@ house	-	4	alder	1935	1220			1.	37	14	67	14	1.
24	*	Thannot	adus	#	1 de dan		d.	, ôtie	tist	1.50			2	Mac	07	lor	4	1
	17	These	Theme	1.	1 Schund	- 14	d	when	19370	1000			Bamon	Guint	#	9 mil	4	
		1	pero -	11		-	+	-	for				qui	Luchane	97.4	7.1.9.	4	fide
		1	- Marcal	1	×-	-	1	-	ge	1.000			Thomas	Junger	21	herdrynn	- 4	methe
	<i>¥</i>	Edgar	Frank	41	ana.		4.		-	1000		+	Mazzini	habert	11	Gran -	1	m. 9:
		12	grapt	11	Agitin	- 7	+		EL MB	1000		4	Sourgie	- Cana	4	Farmelt	- 4	4 20
		1. 14	Tolkin	11	11 12	~	der.		ELMI	1000			10	Guile	92	1.04		-
	+	Jiniste	Hotenat	11	hagen		4.	terr.	22AUL		1	1	game	O Dia	1	5 Hard		inning
	Č.	744012	my	1	8 a	- /	*	and a	-	- Contraction			2	and and	12	12000	- 4	ser.
	7	Radit	- Maria	"	C. C. Comme	-	-	There		-		7		rega	"	-		

Tables obtained after segmentation and dewarping.

1: S. Ares Oliveira, B.Seguin, and F. Kaplan, "dhSegment: A generic deep-learning approach for document segmentation," in Frontiers in Handwriting Recognition (ICFHR), 2018 16th International Conference on, pp. 7-12, IEEE, 2018.

7

Information localization: from tables to table rows

- Detection of the baselines in the table.
- Extraction of a rectangle for each detected baseline, knowing the height of the rows.

DEDUCATION	NEXEDOS File orientere, related harmonic no rust	-		AMER	LIN		MELINA	Clark	Prose los patrons, checks d'orderprise, checks a d'order
	des das da		PRENORS			un		PROFESSION	Front los con-
tantan in sites	antinen eringen uts	-	7				ch d de minogo	1 H	present for the Theory temperature ged for one- phase.
Pine	22 1	Valade	mental		Sind		12	ed	1832
to Amitin	4	Queros	Rend	91	A	ía l	4	1 cining	12174
-	9	Kornlish	missuelar	29	1000	Earley	4	migricant	11374
	4	2 Contucci	mario	11	A ta	là la	4	Lucitan	lezun
a court			Comina	10	A t			1 ans	22403
- August	4	Donato	herrow	96	las me		1 d	As	18394
. 20.00	4	Leconvoisier	Amille	95	Vien		4	alit	10388
	4	Rown	Res	19	@ Anne	and	1	dans	Erren
155.65	54	Inchorn.	Abrille	72	A Li	ue la	d	reiller	2139 a
-	10	a has bianco	I durado	74	A the	le .	. 2	-inine	18374
12574		1	hune	81	A -			1.	home
100000	1	y martiska	a theira	91	A hada	take	1	mainier	1200
Bed das	P . 3	Recher	tilled	52	In		4	miles	21395
Samucines	· _		mathilde	14	or		i	1	risac
1	- 4	, heals	Clanders	22	100		4	me que	here
The second			Buline	82	OP		4	and	Arene
-	32		fland	02	100		4	me da	11374
121	-	1	an	01	100		4	1	16374
15-1-		1	War.	01	100		4	1	12 sm
min	1	Hemon	luis.		Q. m.		d		01300
-		Reis	Smilene.	43	m.1.8.		4	1. Jack	21246
	12 and	Choma	Charge	96	herden		4	1 mile al	u sac
	1	Massini	Interester	11	landes		1	- q.	Jusin 1
- Lineart	4	2 Louris	land	20	Hamilt		4	and Reason	usin
- Deces		1	larise.	91	1. 20		u	1 same	48397
- Traile	9 -	Banne	Bue.	15	18 und.		2	incites	21390
		1	Andre	62	1 de Cuise		ú	sur	
	÷	in	Augle	4	1 since		4	made	-
	-	×	din .	30-		-	2.		ate
1	14	Bres	Turne	41	A que	and	1	any Ore	ham
		0	-		11			-	1. C. S.

Results of baseline detection on a table.

Valade	Wartial	09	Gironde	- the	et anisimise 1832
Guerro	Raout	93	@ Cuba	ch -	inginiare 12174
Howalish	Misegyplow	72	O J. Figue		négriant 18374
Contucci	Maris	89	D Hale	md	requisitan 22 ho
<i>.</i> .	Gonina	87	æ.	m	mus erhor
Donato	Georges	96	144. m. m	d	2. commercial 18374
leconvoisier	Gamille	95	meine		artist 1938 3
Brown	Rose	89	Amérique	m_ch	sens Erter
Eschopp	Shiphile	72	& Luisse .	m ch -	veilleur 2139 9
Capobianco	Eduardo	89	& take	m_ch	requiant 18374
	Anna	89	Ø ø	m	~ 103ju
Martiska	Gatherine	98	D Autriche		cuisiniere 21399

Line images extracted using detected baselines.

8

Data annotation and logical column segmentation

- Encoding of the logical separation into columns using a semantic token "/".
- Avoid the need for column segmentation by teaching the HTR model to predict the logical separation into columns beside predicting the handwritten text.

Wartial Valade / martial / 09 / Gironde / ¤ / ¤ / ch / ¤ / ap. cuisinier / 18371 93 Cuervo / Raoul / 93 / Cuba / ¤ / ¤ / ch / ¤ / ingénieur / 12174

Examples of labels.



Example of table where words are overlapping columns.

Dataset	Train # of lines	Validation $\#$ of lines	Test # of lines	# of writers
POPP (Generic)	$3840 \ (128 \text{ pages})$	$480 \ (16 \text{ pages})$	$480 \; (16 \; pages)$	80
Belleville	1140 (38 pages)	150 (5 pages)	$180 \ (6 \text{ pages})$	1
Chaussée d'Antin	625	78	77	10

Details about the POPP datasets.

Handwriting recognition



Diagram of the POPP processing pipeline.

Handwriting recognition: supervised learning

- Architecture:
 - Line HTR architecture from our team at LITIS described in Coquenet-2022².
- Recognition results:
 - Generic dataset: Not as good as the results obtained with the same architecture on IAM or RIMES, this dataset seems more challenging.
 - Belleville: shows the improvement that can be expected from writer specialization.

Dataset	Val CER	Val WER	Test CER	Test WER
Generic (multi-writer)	6.86~%	18.66~%	7.08~%	$19,\!05~\%$
Belleville (writer-specific)	3.36~%	7.47~%	3.65~%	8.65~%

Recognition results obtained on the POPP datasets using supervised learning.

2: D. Coquenet, C. Chatelain and T. Paquet. "End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network." *IEEE transactions on pattern* **11** *analysis and machine intelligence* PP (2022)

Handwriting recognition: self-training

- Method inspired by Q. Xie's 2019³, which combines self-study and noisy student techniques.
- Useful when a large amount of unlabeled data is available.
- In our case, we use 2.4 million line images selected randomly from the 1926 census.



3: Q. Xie, M. -T. Luong, E. Hovy and Q. V. Le, "Self-Training With Noisy Student Improves ImageNet Classification," 2020 IEEE/CVF Conference on Computer Vision **12** and Pattern Recognition (CVPR), 2020, pp. 10684-10695

Handwriting recognition: model architectures

During our experiments, 3 architectures were used:

- Architecture A: the line HTR architecture described in Coquenet-2022.
- Architecture B: Architecture A with scaling factor of 1.5 for depth and 1.25 for width.
- Architecture C: Encoder of architecture B with a BLSTM-based decoder.



Handwriting recognition: results

- A more powerful architecture can further improve the results when a self-training iteration bring no improvement.
- The addition of an LSTM part allows, compared to a fully convolutional model, to implicitly learn a language model from the large volume of data encountered.

Model	Architecture	Dataset	CER (%) test	WER (%) test
Initial (Model 0)	A	Generic	7.08	19.05
Student 1	A	Generic	6.12	17.12
Student 2	А	Generic	5.97	16.83
Student 2bis	А	Generic	6.02	16.89
Student 3	В	Generic	5.43	15.50
Student 4	С	Generic	4.52	13.57
Student 4 specialized	С	Mono-writer	2.66	6.37

Grammar and post-processing steps



Diagram of the POPP processing pipeline.

Grammar and post-processing steps

- The output of the recognition model is processed by grammars written in Thrax, compiled into Weighted Finite State Transducers⁴ and decoded by Kaldi.
- We have created SIGRA, a Python Framework that facilitates the use of grammars for handwriting recognition by linking Thrax and Kaldi.
- The grammars apply a rejection process for each cell and normalize the accepted fields.



4: M. Mohri, F. Pereira, and M. Riley. 2002. Weighted finite-state transducers in speech recognition. Computer Speech & Language 16, 1 (2002), 69–88

Grammar and post-processing steps



Cuervo / Raoul / 93 / Cuba / ¤ / ¤ / ch / ¤ / ingénieur / 12174

Prediction of the model.

CUERVO	RAOUL	Н	1893			CUBA	CUBA	С	СН		INGENIEUR	12174
VALADE	MARTIAL	Н	1909		GIRONDE	FRANCE	FRANCAISE	С	СН	APP	APPRENTI CUISINIER	18371
Noms	Prenom	Sexe	Annee de naissance	Ville de naissance	Departement de naissance	Pays de naissance	Nationalite	Etat matrimonial	Situation par rapport au chef de menage	Statut profession	Profession	Code metier

Output of the grammar and post-processing steps.

Conclusion

- The POPP datasets⁵
 - We publish our annotated datasets containing ground truth for handwriting recognition and line coordinates.

Dataset	Train # of lines	Validation # of lines	Test # of lines	# of writers
POPP (Generic)	$3840 \ (128 \text{ pages})$	480 (16 pages)	480 (16 pages)	80
Belleville	1140 (38 pages)	150 (5 pages)	180 (6 pages)	1
Chaussée d'Antin	625	78	77	10

Details about the POPP datasets.

- SImple GRAmmar toolkit (SIGRA)⁶
 - We open source the code of SIGRA, a Python framework that facilitates the use of WFSTs with **Kaldi** for handwriting recognition.



6: https://gitlab.com/projet-popp/sigra





QR code for SIGRA



Conclusion

• POPP Project

- Our pipeline processed the complete census of 1926, 1931, 1936 with a total of 300k pages, 9 million lines and several hundred writers.
- The POPP database is currently being exploited by a team composed of historians, sociologists, and economists.

• Perspectives

- The processing pipeline could be adapted easily for the other French population census of the time period because the same census procedures and the same table templates were used.
- The self-training method could be further explored regarding handwriting recognition.

Bibliography

- S. Ares Oliveira, B.Seguin, and F. Kaplan, "dhSegment: A generic deep-learning approach for document segmentation," in Frontiers in Handwriting Recognition (ICFHR), 2018 16th International Conference on, pp. 7-12, IEEE, 2018.
- D. Coquenet, C. Chatelain and T. Paquet. "End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network." IEEE transactions on pattern analysis and machine intelligence PP (2022)
- Q. Xie, M. -T. Luong, E. Hovy and Q. V. Le, "Self-Training With Noisy Student Improves ImageNet Classification," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10684-10695
- M. Mohri, F. Pereira, and M. Riley. 2002. Weighted finite-state transducers in speech recognition. Computer Speech & Language 16, 1 (2002), 69–88

Writer specialization combined with self-training

- 1. A model is initialized with the weights of the best student model
- 2. The model is trained on the Belleville dataset (mono-writer)
- 3. The model perform inference on every unlabeled data of the Belleville district
- 4. A second model is trained on the generated pseudo-labels