



EXO-POPP Project

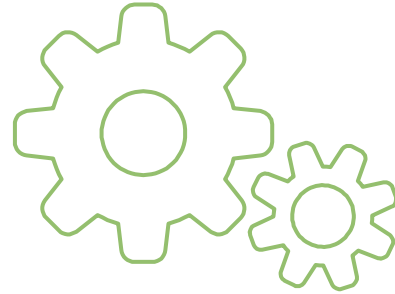
Extraction of named entities from marriage records

Thomas Constum, Pierrick Tranouez, Thierry Paquet

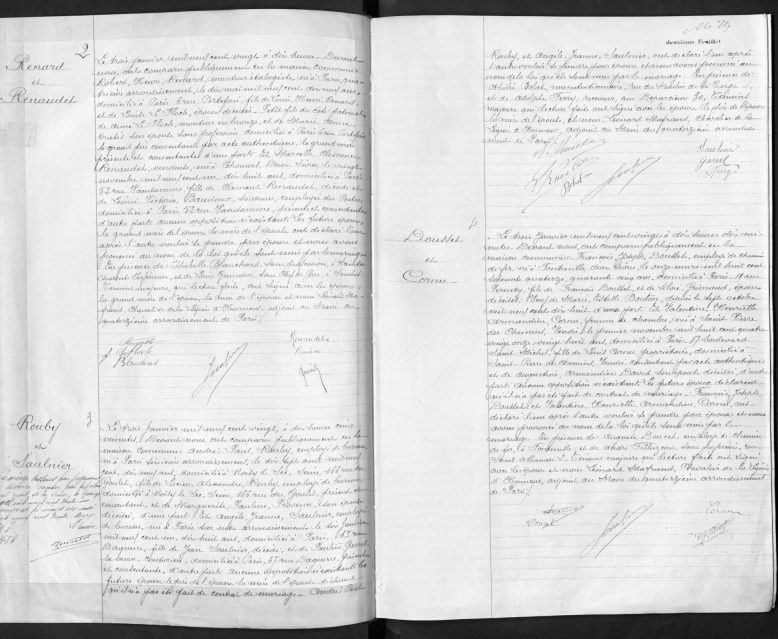
1



Presentation of the project



1.1 The Exo-POPP project

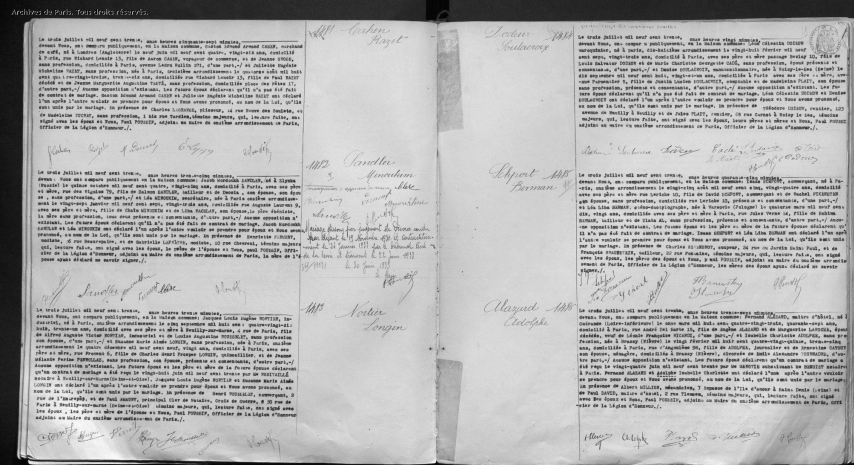
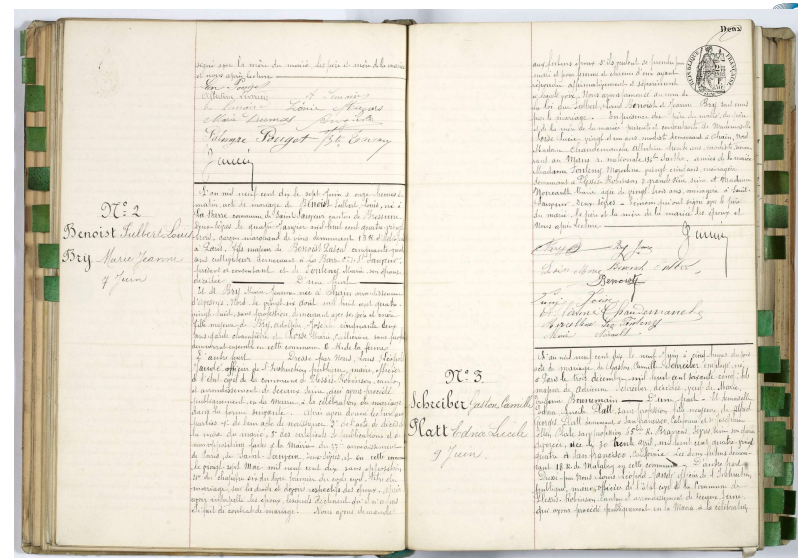


Date mariage	Prénom mari	Nom mari	Profession mari	Lieu naissance mari	Date naissance mari	...
03/01/1920	Robert	Renard	monteur horlogiste	Paris	10/05/1900	...
03/01/1920	André	Rouby	employé de bureau	Paris	17/08/1898	...
03/01/1920	François	Doussier	employé de chemin de fer	Paris	05/01/1902	...

Objective: build a database from the extraction of named entities from 150 000 marriage records

1.2 EXO-POPP Corpus

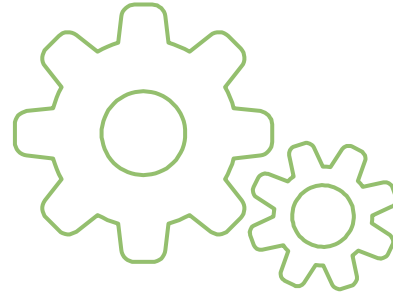
- Marriage records from the former department of the Seine, now divided between the Hauts-de-Seine, Paris, Seine-Saint-Denis and Val-de-Marne
- Manuscript: period 1880-1920 (about 80% of the data)
- Tapuscript: period 1930-1940
- Variability in:
 - the size of the booklet and the margins
 - writing style
 - the wording of the acts
 - the scanning conditions



2



Handwriting recognition

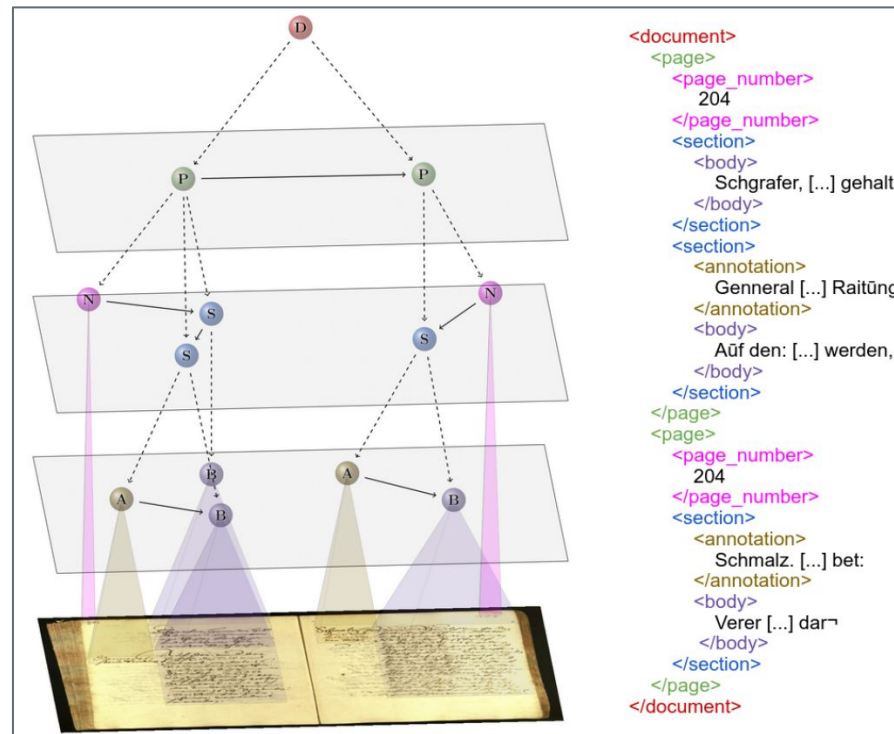


2.1

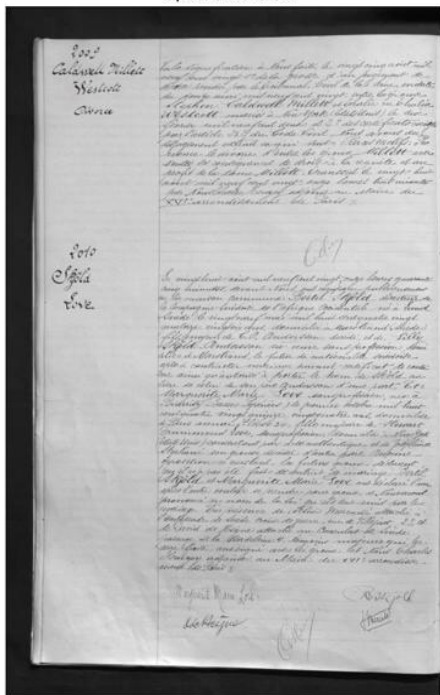
Handwriting recognition with the DAN



- DAN¹ (Document Attention Network)
 - handwriting recognition on a complete document of one or two pages without layout annotation
 - semantic recognition of text zones
- Objective: apply DAN to the EXO-POPP corpus to remove the text field detection step



Input document



cer: 4.72% loer: 0.0% map cer: 84.29%

Iteration: 2

<mariage>

Ground truth

0 <page>
1 <mariage>
2 <mariage-noms>
3 2009
4 Caldwell Millett
5 Westcott
6 Divorce
7 <mariage-noms>
8 <paragraphe>
9 Vu la signification à Nous faite le vingt-cinq août mil
10 neuf cent vingt 1^{er} de la grosse d'un jugement de
11 divorce rendu par le Tribunal Civil de la Seine, en date
12 du douze mai mil neuf cent vingt entre les époux
13 Stephen Caldwell Millett et Chalie ou Chalia
14 Westcott mariés à New-York (Etats Unis) le dix
15 février mil neuf cent deux et 2^e des certificats exigés
16 par l'article 252 du Code Civil, Nous avons du
17 dit jugement extrait ce qui suit: Par ces motifs: Pro
18 nonce le divorce d'entre les époux Millett avec
19 toutes ses conséquences de droit à la requête et au
20 profit de la dame Millett, transcrit le vingt huit
21 août mil neuf cent vingt: onze heures trois minutes
22 par Nous Charlie Heuzey adjoint au Maire du
23 XVI arrondissement de Paris./.
24 </paragraphe>
25 <mariage>
26 <mariage>
27 <mariage-noms>
28 2010
29 Shyld
30 Love
31 <mariage-noms>
32 <paragraphe>
33 Le vingt huit août mil neuf cent vingt, onze heures quarante
34 cinq minutes, devant Nous ont comparu publiquement
35 en la maison commune: Bertil Igöld directeur de
36 la Compagnie Suédoise de l'Afrique Occidentale né à Lund
37 (Suède) le vingt neuf mai mil huit cent quatre vingt
38 quatorze, vintsix ans, domicilié à Mastrand (Suède)
39 fils majeur de C.O. Andersson, décédé, et de Lily
40 Sköld Andersson sa veuve sans profession, domi
41 ciliée à Mastrand, le futur de nationalité suédoise
42 apté à contracter mariage suivant certificat de coutu
43 me, ainsi qu'autorisé à porter le nom de Sköld au
44 lieu de celui de son père Andersson d'une part. Et:
45 Marguerite Marie Love sansprofession, née à
46 Biarritz (Basses-Pyrénées) le premier octobre mil huit
47 cent quatre vingt quinze vingt quatre ans, domiciliée
48 à Paris avenue Kléber 54, fille majeure de Stenvar
49 Drummond Love, sansprofession, domicilié à New-York
50 (Etats Unis) consentant par acte authentique et de Joséphine
51 Séphani son épouse décédée d'autre part. Aucune
52 opposition n'existant, les futurs époux déclarent
53 qu'il n'a pas été fait de contrat de mariage. Bertil
54 Sköld et Marguerite Marie Love ont déclaré l'un
55 après l'autre vouloir se prendre pour époux et Nous avons
56 prononcé au nom de la loi qu'ils sont unis par le
57 mariage. En présence de: Felix Mercadé attaché à
58 l'Ambassade de Suède croix de guerre, rue de Villejust 23, et
59 de David de Høyne attaché au Consulat de Suède
60 passage de la Madeleine 4 témoins majeurs qui lec
61 ture faite, ont signé avec les époux et Nous Charles
62 Heuzey adjoint au Maire du XVI arrondisse
63 ment de Paris./.
64 </paragraphe>

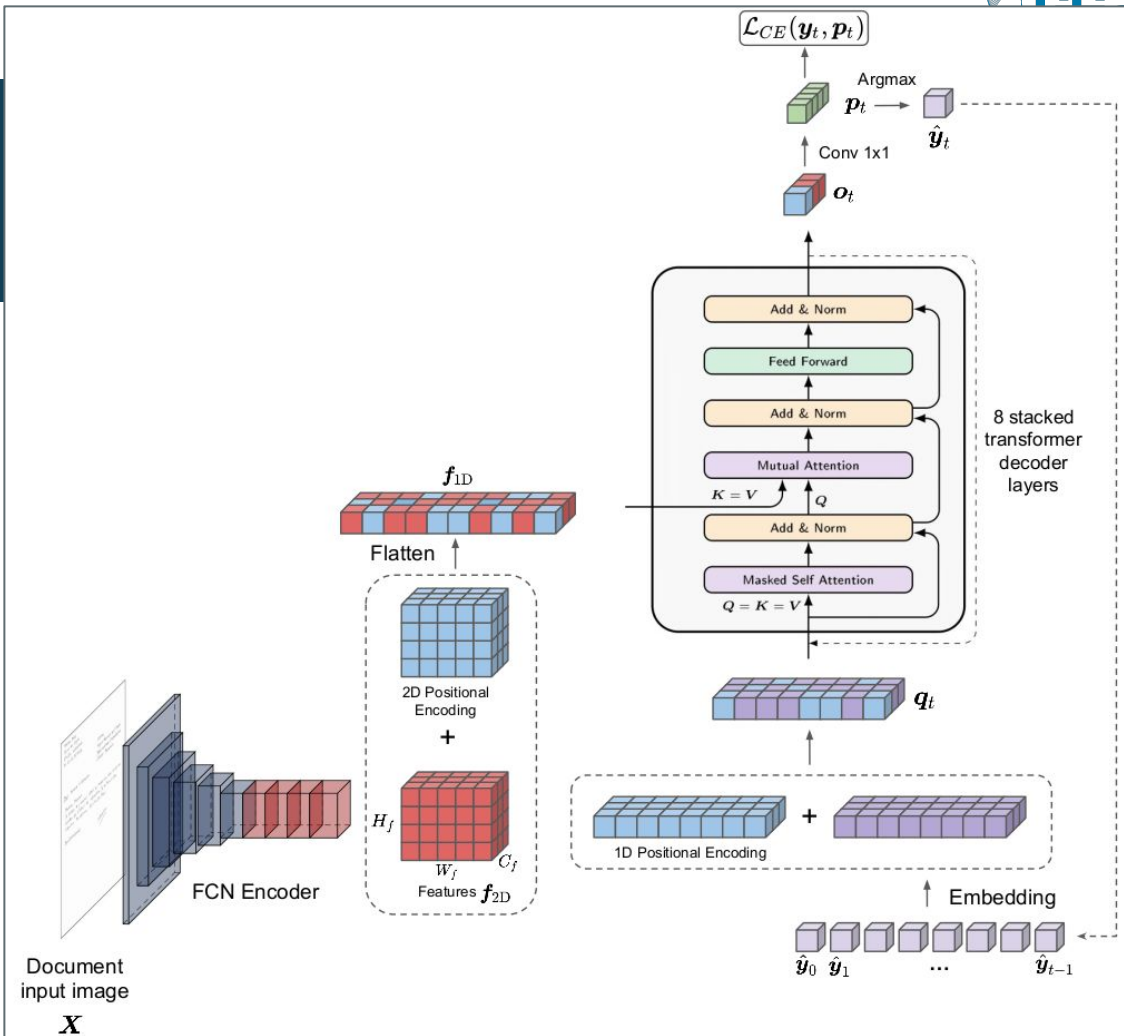
Prediction

0 <page>
1 <mariage>
2

2.1

Handwriting recognition with the DAN

- Encoder: same architecture as the VAN (Coquenot2022)
- 2D positional embedding of the feature map (Singh2021)
- Decoder: variant of the Transformer architecture (Vaswani2017)



2.1

Handwriting recognition with the DAN



Training process:

- Pre-training of the encoder on synthetic lines
- Training of the DAN on synthetic documents with an increasing number of lines (from 1 to max_lines) and 90% chance of generating synthetic data
- Continued training with a progressive decrease in the proportion of synthetic data until 20% chance of generating synthetic data is reached

Annotated data

	Train	Valid	Test
# simple pages	254	24	24

611 si vous proposiez des services d'assurance

296 **est des plus gênant pour lire la presse du jour au matin.**
Veuillez agréer, Madame, Monsieur, mes

270 **BMMPS49, je viens d'acquérir une automobile et je**
le 24 juillet 1984 Veuillez agréer mes salutations distinguées
le volume sonore augmente. Ceci est très désagréable d'autant

N° 395 *Objet : demande de mensualisation de l'impôt sur le revenu*
je souhaiterais connaître vos différents produits
-ne reçu et attend une réponse de votre part.
banque. Voici mes nouvelles coordonnées bancaires.
Objet : demande de mensualisation de l'impôt sur le revenu

N°219
14 août

Objet : Demande de renseignements concernant l'ouverture
ranchise que j'ai souscrite chez vous. Je vous explique : la semaine
Naif assurances - Le Bourg - 15240 Saignes
de bénéficier, pour un montant d'une commande de 20 euros,
que j'ai oublié de commander 3 paires de chaussettes, référence

2.1

Handwriting recognition with the DAN



Training process:

- Pre-training of the encoder on synthetic lines
- Training of the DAN on synthetic documents with an increasing number of lines (from 1 to max_lines) and 90% chance of generating synthetic data
- Continued training with a progressive decrease in the proportion of synthetic data until 20% chance of generating synthetic data is reached

Annotated data

	Train	Valid	Test
# simple pages	254	24	24

*vous demander un échelonnement de mes paiements
ai envoyé le mois dernier et qui est resté sans
Suite à la tempête du 01/01/99 mon toit a été
5 octobre 2006, j'ai constaté une anomalie dans le
Je vous écris pour vous exprimer mon indignation. Non
vous demander un échelonnement de mes paiements
Objet : Augmentation (x2) des quantités de CD vierges
afin d'augmenter la quantité de CD vierges que j'ai demandés,
les plus brefs délais. Or à ce jour je n'ai
d'agréer, Monsieur, mes salutations distinguées
Objet : Demande de documentations concernant les contrats
"tapage" nocturne est de votre ressort et je*

443

**43123 Cordelle
Madame, Monsieur,**

Je me tiens à votre disposition au 01.12.65.65.11
Je vous écris pour vous exprimer mon mécontentement.

**de mes versements sur mon compte
cause d'une erreur sur l'interprétation de mes**
Objet : Demande de rendez-vous avec conseiller
de bien vouloir me faire parvenir dans les

l'écoute de mon portable, pour un moment

Je me tiens à votre disposition au 01.12.65.65.11

Veillez agréer, Monsieur, l'expression de mes

merveilleuse commande. Je souhaite recevoir, incessamment

Objet : demande de mensualisation de l'impôt sur le revenu
Trésorier, l'expression de mes sentiments

Je sollicite par la présente une remise d'impôt.

agréer mes salutations les plus distinguées

à mon domicile dans les meilleurs délais.

frein provoquant alors une collision avec le

Je désire fermer mon compte KKWQY71 concernant une

D'autre part, je tiens à vous faire remarquer que le montant de l'assurance

et je vous prie d'agréer, Madame, Monsieur, l'expression

Afin de faire mon choix, pourriez-vous me faire

la situation actuelle faite donc que je n'aurai plus besoin

Objet : demande de mensualisation de l'impôt sur le revenu
Comme demandé lors de notre conversation téléphonique du

Objet : Diminution des versements sur plan d'épargne logement

Par ce courrier, je voudrais vous faire part

j'ai rencontré la voiture du facteur

Vous en remerciant d'avance, je vous prie d'agréer,

Objet : résiliation d'assurance habitation (ref NRDGU69)

unréclamation. Depuis que la rue de la Papeterie est

Cependant je vous informe qu'il ne fonctionne pas.

factures que vous m'avez présentées je sollicite de

Dans l'attente d'une réponse rapide, je vous
votre entreprise et j'aimerais la modifier ; je voudrais

2.1

Reconnaissance d' écriture Présentation du DAN



Utilisation de documents synthétiques

- Permet au modèle de converger en s'entraînant sur un grand volume de données
- Permet d'augmenter progressivement le nombre de lignes (curriculum learning)
- Inconvénient: besoin de développer une méthode de génération de documents synthétiques similaires à nos données mais avec suffisamment de diversité

Objet : accident de la route
58 rue de la République
15 rue Principale
A Desseling, le 25 octobre 2006
Le Bourg
58 rue de la République

vous demander un échelonnement de mes paiements
ai envoyé le mois dernier et qui est resté sans
Suite à la tempête du 01/01/99 mon toit a été
5 octobre 2006, j'ai constaté une anomalie dans le
Je vous écris pour vous exprimer mon indignation. Non
vous demander un échelonnement de mes paiements
Objet : Augmentation (x2) des quantités de CD vierges
vous parvenez Commande de 2 lots de 10 unités de CD vierges
afin d'augmenter la quantité de CD vierges que j'ai demandés.
les plus brefs délais. Or à ce jour je n'ai
d'agréer, Monsieur, mes salutations distinguées
Objet : Demande de documentations concernant les contrats
"tapage" nocturne est de votre ressort et je

443

43123 Cordelle
Madame, Monsieur,

Je me tiens à votre disposition au 01.12.65.65.11
Je vous écris pour vous exprimer mon mécontentement.
de mes versements sur mon compte
cause d'une erreur sur l'interprétation de mes
Objet : Demande de rendez-vous avec conseiller
de bien vouloir me faire parvenir dans les
l'écoute de mon portable, pour un moment
Je me tiens à votre disposition au 01.12.65.65.11
Veuillez agréer, Monsieur, l'expression de mes
merveilleuse commande. Je souhaite recevoir, incessamment
Objet : demande de mensualisation de l'impôt sur le revenu
Trésorier, l'expression de mes sentiments
Je sollicite par la présente une remise d'impôt.
agréer mes salutations les plus distinguées
à mon domicile dans les meilleurs délais.
frein provoquant alors une collision avec le
Je désire Fermer mon compte KKWQY71 concernant une
l'autre part, je tiens à vous faire remarquer que le montant de l'assurance
et je vous prie d'agréer, Madame, Monsieur, l'expression
Afin de faire mon choix, pourriez-vous me faire
Ma situation actuelle fait donc que je n'aurais plus besoin
Objet : demande de mensualisation de l'impôt sur le revenu
Comme demandé lors de notre conversation téléphonique du
Objet : Diminution des versements sur plan d'épargne logement
Par ce courrier, je voudrais vous faire part
j'ai rencontré la voiture du facteur
Vous en remerciant d'avance, je vous prie d'agréer,
Objet : résiliation d'assurance habitation (ref NRDGU69)
unréclamation. Depuis que la rue de la Papeterie est
Pendant je vous informe qu'il ne fonctionne pas.
factures que vous m'avez présentées je sollicite de
Dans l'attente d'une réponse rapide, je vous
votre entreprise et j'aimerais la modifier ; je voudrais

2.2

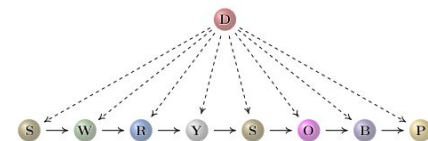
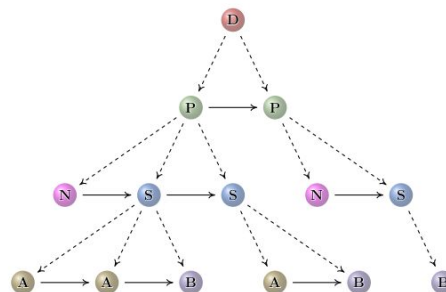
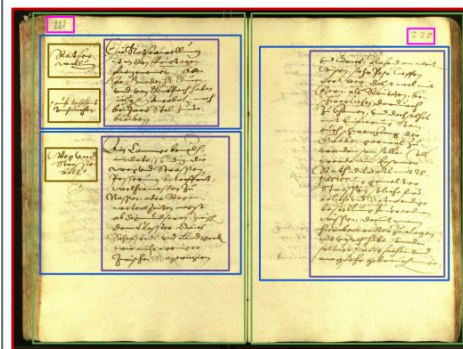
Handwriting recognition with the DAN



Metrics

- Joint recognition of layout and text requires the use of new metrics
- LOER (Layout Ordering Error Rate): evaluation of structure recognition and reading order

$$\text{LOER} = \frac{\sum_{i=1}^K \text{GED}(\mathbf{y}_i^{\text{graph}}, \hat{\mathbf{y}}_i^{\text{graph}})}{\sum_{i=1}^K n_{e_i} + n_{n_i}}$$



2.2

Handwriting
recognition with the
DAN

Metrics

- MAP_{CER} : Mean Average Precision using CER as a threshold. Joint evaluation of layout and handwriting recognition.

$$\text{AP}_{\text{CER}_c}^{5:50:5} = \frac{1}{10} \sum_{k=1}^{10} \text{AP}_{\text{CER}_c}^{5k}.$$

Example

- Prediction: "<X>text1</X><A>text2<A>text3"
- GT: "<X>text1</X><A>text4<A>text2"

Score	Prediction
X	
80%	text1
A	
84%	text2
80%	text3
B	
85%	text2text3

	Ground truth
X	
	text1
A	
	text4
	text2
B	
	text4text2

2.2

Reconnaissance d'
écriture manuscrite
Résultats du DAN

Métriques

- MAP_{CER} : Mean Average Precision utilisant le CER comme seuil. Évaluation conjointe du layout et de la reconnaissance d'écriture.

$$\text{AP}_{\text{CER}_c}^{5:50:5} = \frac{1}{10} \sum_{k=1}^{10} \text{AP}_{\text{CER}_c}^{5k}.$$

$$\text{mAP}_{\text{CER}} = \frac{\sum_{c \in S} \text{AP}_{\text{CER}_c}^{5:50:5} \cdot \text{len}_c}{\sum_{c \in S} \text{len}_c}$$

2.2

Handwriting
recognition
Results of the DAN

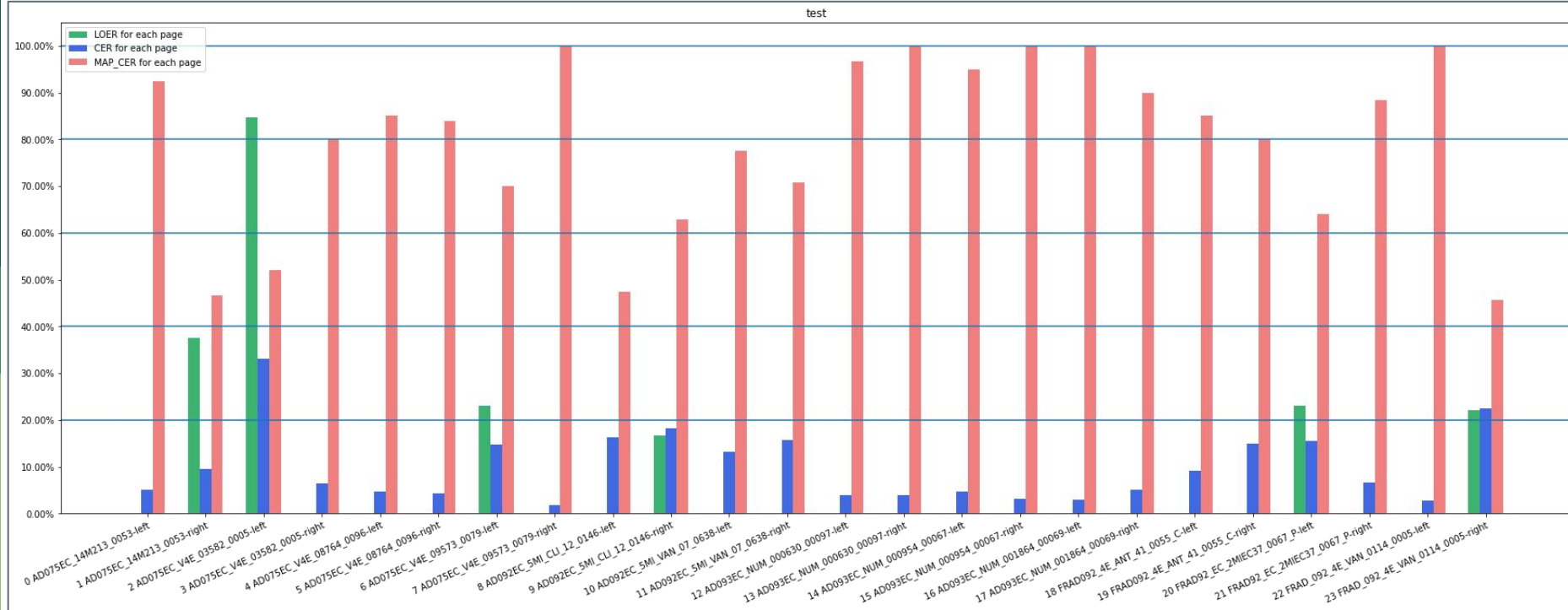
Results on Exo-POPP

	CER	WER	LOER	MAP_CER
Validation	8.01%	17.4%	8.97%	78.6%
Test	10.12%	20.32%	10.10%	79.5%

Results on other datasets

	CER	WER	LOER	MAP_CER
READ Test - simple pages	3.53%	13.33%	5.94%	92.57%
RIMES Test	4.54%	11.85%	3.82%	93.74%

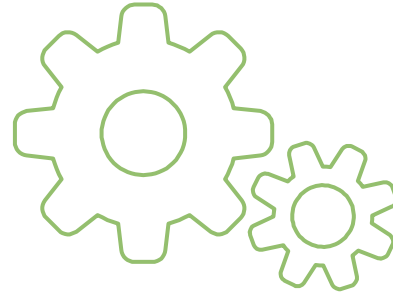
2.2

Handwriting
recognition
Results of the DAN

3



Named entity extraction



3.1 | Methodology

Recognition of named entities in handwritten documents

- Objective: Design an architecture that jointly performs handwriting recognition and named entity recognition
- In total: 137 types of tags. Examples: "Department of residence of the wife's father", "Child's first name".

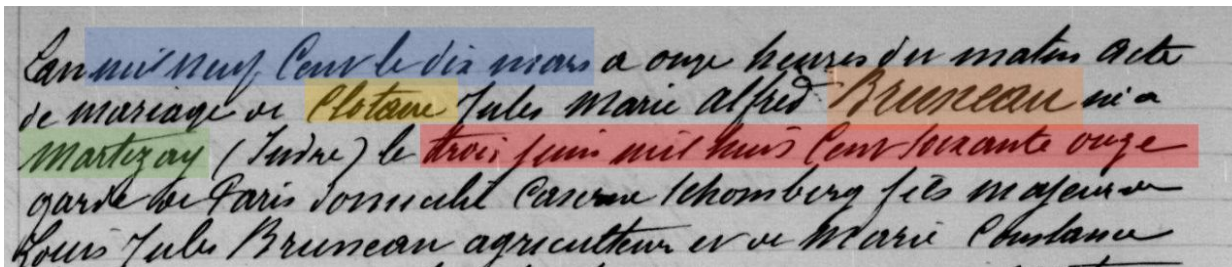
Date : 10 Mars 1900

Prénom du mari : Clotaire

Nom de famille du mari : Bruneau

Lieu de naissance du mari : Martizay

Date de naissance du mari: 3 juin mille huit cent soixante onze



Le sous-mari Clotaire le dix mars a onze heures du matin acte
de mariage de Clotaire Jules Marie Alfred Bruneau né à
Martizay (Indre) le trois juin mil huit cent soixante onze
garde de Paris domicilié Casseau Schomburg fils m. a. p. m.
Louis Jules Bruneau, agriculteur et de Marie Claudine

3.1

Named entity
recognition

Étiquettes

Rechercher

JEUX SÉLECTIONNÉS JEUX UTILISÉS ÉT >

Exo-popp

- Administratif
- Mari
- Epouse
- Témoin
- Enfant
- Père
- Mère
- Ex-epoux
- Naissance
- Résidence
- Décès
- Divorce
- Majeur
- Sexe
- Registre
- Contrat de mariage
- Oubli en marge
- Lien
- Age
- Divorce 1ere noce
- Legitime
- Appel aux armes
- Divorce (Transcription)
- Divorce (Mention)
- Date
- Adresse
- Profession
- Prenom
- Nom
- Veuf
- Consentant
- Non consentant
- Absent
- Disparu
- Non dénommé
- Tribunal
- Pays
- Département
- Ville
- Numéro voie
- Type voie
- Nom voie
- Année
- Mois
- Jour
- Heure
- Minute
- Présent

Le vingt neuf mai mil neuf cent vingt a dix heures cinquante devant Nous, ont comparu publiquement en la maison commune: François Santoni, comptable né à Serra-di-Fiumorbo (corse) le premier octobre mil huit cent quatre vingt quatre, trente cinq ans, domicilié à Paris 26 rue Duméril, fils de Simon Santoni et de Marie Hélène Grudicelli, époux décédés, d'une part et Eugénie Emilie Pons, couturière, née à Maurs (cantal) le vingt deux Janvier Mil huit cent quatre vingt dix sept vingt trois ans, domiciliée à Paris, 41 rue de Saintonge, fille de Pierre Henri Pons, horloger, et de Jeanne Anaïs Maynard, ménagère, époux domiciliés 41 rue de Saintonge, présents et consentants. d'autre part, aucune opposition n'existant, Le futur époux la future épouse ses père et mère, ont déclaré qu'il

3.1

Named entity
recognition

Étiquettes

Rechercher

JEUX SÉLECTIONNÉS JEUX UTILISÉS ÉT >

Exo-popp

Administratif Mari Epouse Témoin Enfant

Père Mère Ex-époux Naissance Résidence

Décès Divorce Majeur Sexe Regist

Contrat de mariage Oubli en marge Lien Age

Divorce 1ere noce Legitime Appel aux armes

Divorce (Transcription) Divorce (Mention) Date

Adresse Profession Prenom Nom Veuf

Consentant Non consentant Absent Disparu

Non dénommé Tribunal Pays Département

Ville Numéro voie Type voie Nom voie Année

Mois Jour Heure Minute Présent

Le vingt neuf mai mil neuf cent vingt a dix heures cinquante devant Nous, ont comparu
publiquement en la maison commune: François Santoni, comptable né à
Serra-di-Fiumorbo (corse) le premier octobre mil huit cent quatre vingt quatre, trente cinq
ans, domicilié à Paris 26 rue Duméril, fils de Simon Santoni et de Marie Hélène Grudicelli,
époux décédés, d'une part et Eugénie Emilie Pons, couturière, née à Maurs (cantal) le vingt
deux Janvier Mil huit cent quatre vingt dix sept vingt trois ans, domiciliée à Paris, 41 rue de
Saintonge, fille de Pierre Henri Pons, horloger, et de Jeanne Anaïs Maynard, ménagère,
époux domiciliés 41 rue de Saintonge, présents et consentants. d'autre part, aucune
opposition n'existant, Le futur époux la future épouse ses père et mère, ont déclaré qu'il

3.1

Named entity
recognition

Using DAN for joint recognition of layout, handwriting and named entities

Encoding of named entities in the DAN charset:

“(m)(n) Santoni et Pons(N)(b)Le 17 17 vingt neuf 17 17 mai 17 17 mil neuf cent vingt a 17 dix heures cinquante devant Nous, ont comparu publiquement en la maison commune: François Santoni, comptable né à Serra-di-Fiumorbo (corse) le 17 17 premier octobre 17 17 mil huit cent quatre vingt quatre, trente cinq ans, domicilié à Paris 10 26 rue ABCD Duméril, fils de Simon Santoni et de Marie Hélène Grudicelli, époux décédés, d'une part et Eugénie Emilie Pons, couturière, née à Maurs (cantal) le 17 17 vingt deux 17 17 Janvier 17 17 Mil huit cent quatre vingt dix sept vingt trois ans, domiciliée à Paris, 41 rue de Saintonge(B)(M)”

4



Conclusion



4

Conclusion

**Results**

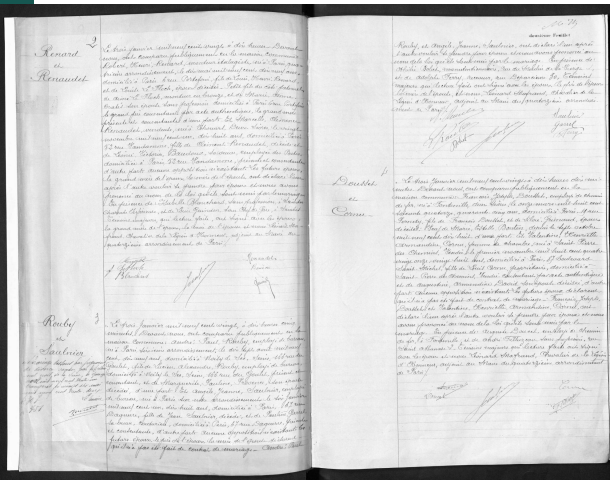
Successful use of DAN on Exo-POPP data for joint layout and handwriting recognition

Perspectives

- Improved performance of HTR
- Using DAN to perform layout recognition, handwriting recognition and named entity extraction together
- Towards a DAN for document understanding?

4

Conclusion



Date mariage	Prénom mari	Nom mari	Profession mari	Lieu naissance mari	Date naissance mari	...
03/01/1920	Robert	Renard	monteur horlogiste	Paris	10/05/1900	...
03/01/1920	André	Rouby	employé de bureau	Paris	17/08/1900	...
03/01/1920	François	Dousset	employé de chemin de fer			...



EXO-POPP project

Friday, October 14, 2022

Thank you for your
attention

-

Any questions?



5.1

Bibliography



- D. Coquenat, C. Chatelain, T. Paquet: “DAN: a Segmentation-free Document Attention Network for Handwritten Document Recognition”
- D. Coquenat, C. Chatelain, T. Paquet: “End-to-end handwritten paragraph text recognition using a vertical attention network,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
- S. S. Singh and S. Karayev, “Full page handwriting recognition via image to sequence extraction,” in 16th International Conference on Document Analysis and Recognition, ICDAR, 2021
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Annual Conference on Neural Information Processing Systems, 2017, pp. 5998–6008.

5.1

Annexes
Résultats du DAN sur
d'autres datasets

Résultats

RIMES

	CER	WER	LOER	MAP_CER
Test - Pages simples	3.53%	13.33%	5.94%	92.57%
Test - Pages doubles	3.69%	14.20%	4.60%	93.92%

READ2016

	CER	WER	LOER	MAP_CER
Test	4.54%	11.85%	3.82%	93.74%

5.2

Reconnaissance d'entités nommées



Niveau	Tags					
1	Administratif	Mari	Epouse	Témoin	Enfant	
1.1	Père	Mère	Ex-Epoux			
2	Naissance	Residence	Deces	Divorce	Majeur	Sexe
	Registre	Contrat de mariage	Oubli en marge	Lien	Age	Divorce 1ere noce
	Legitime	Appel aux armes				
	2.1	Divorce (Transcription)	Divorce (Mention)			
3	Date	Adresse	Profession	Prenom	Nom	Veuf
	Consentement (présent et consentant)	Consentement (décès)	Consentement (absent mais consentant)	Consentement (disparu)	Consentement (non dénommé)	Tribunal
	4	Pays	Département	Ville	Numéro voie	Type voie
	Année	Mois	Jour	Heure	Minute	

3.3

Reconnaissance
d'entités nommées

Essais d'utilisation du DAN pour la NER

Encodage des EN dans le charset du DAN:

Le 17 vingt 17 neuf 17 17 mai 17 17 mil 17 17 neuf 17 17 cent 17 17 vingt 17 a
 17 dix 17 heures 17 cinquante devant Nous, ont comparu publiquement en la maison
commune: François Santoni, comptable né à Serra-di-Fiumorbo
(corse) le 17 17 premier 17 17 octobre 17 17 mil 17 17 huit 17 17 cent 17 17
quatre 17 17 vingt 17 17 quatre, trente cinq ans, domicilié à Paris
 26 rue Duméril, fils de Simon Santoni et de Marie
 Hélène Grudicelli, époux décédés, d'une part et Eugénie Emilie Pons,
 couturière, née à Maurs (cantal) le 17 17 vingt 17 17 deux 17 17
Janvier 17 17 Mil 17 17 huit 17 17 cent 17 17 quatre 17 17 vingt 17 17 dix 17
 17 sept vingt trois ans, domiciliée à Paris, 41 rue
 de Saintonge, fille de