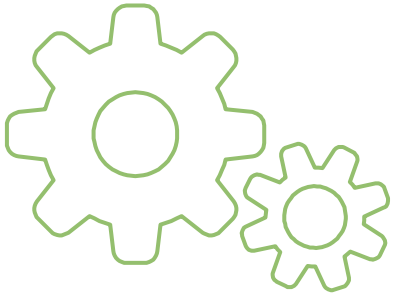
EXO-POPP Project
Extraction of named entities from marriage records

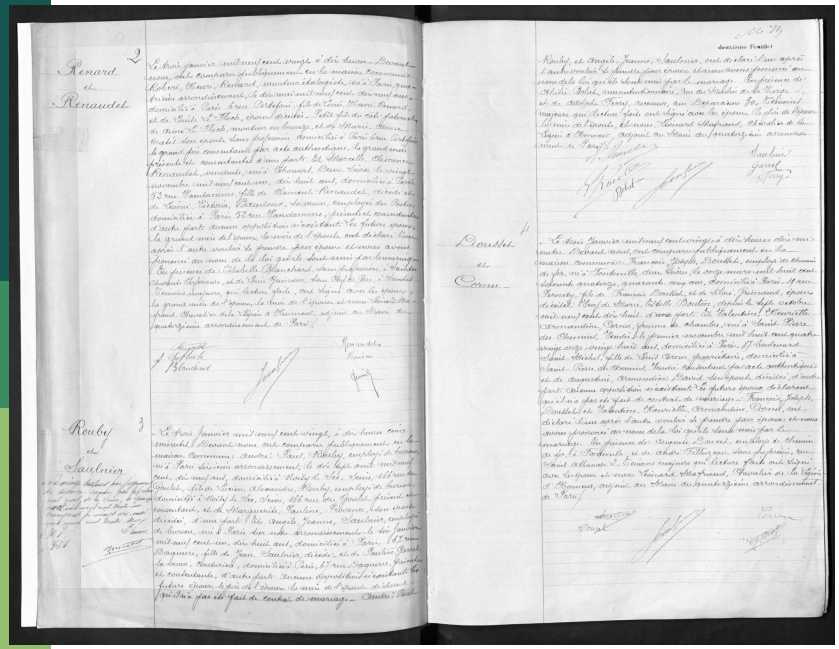
Thomas Constum, Pierrick Tranouez, Thierry Paquet

1

Presentation of the project



1.1 The Exo-POPP project

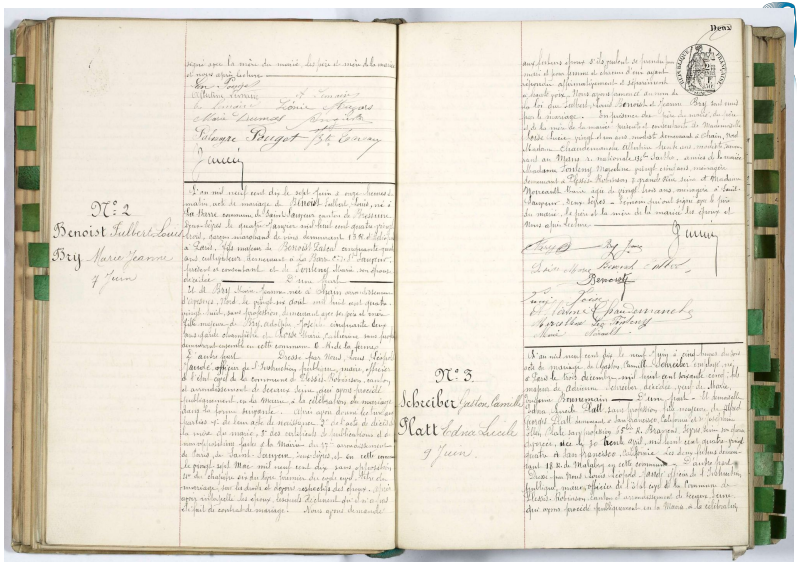


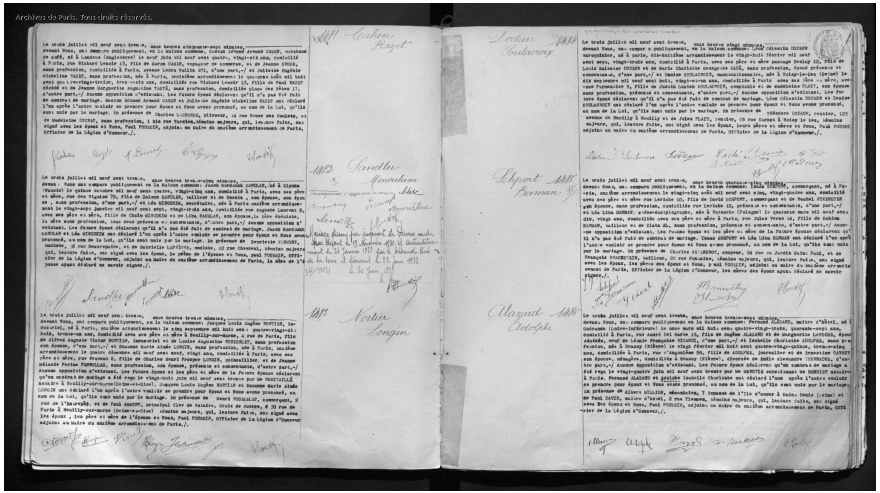
Date mariage	Prénom mari	Nom mari	Profession mari	Lieu naissance mari	Date naissance mari	...
03/01/1920	Robert	Renard	monteur horlogiste	Paris	10/05/1900	...
03/01/1920	André	Rouby	employé de bureau	Paris	17/08/1898	...
03/01/1920	François	Dousset	employé de chemin de fer	Paris	05/01/1902	...

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

1.2 EXO-POPP Corpus

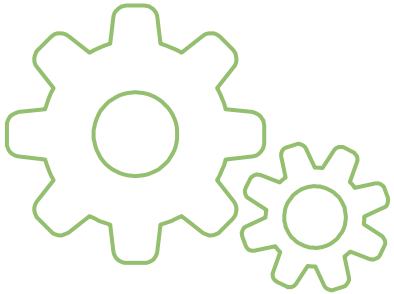
- Marriage records from the former department of the Seine, now divided between the Hauts-de-Seine, Paris, Seine-Saint-Denis and Val-de-Marne
- Manuscript: period 1880-1920 (about 80% of the data)
- Tapuscript: period 1930-1940
- Variability in:
  - the size of the booklet and the margins
  - writing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





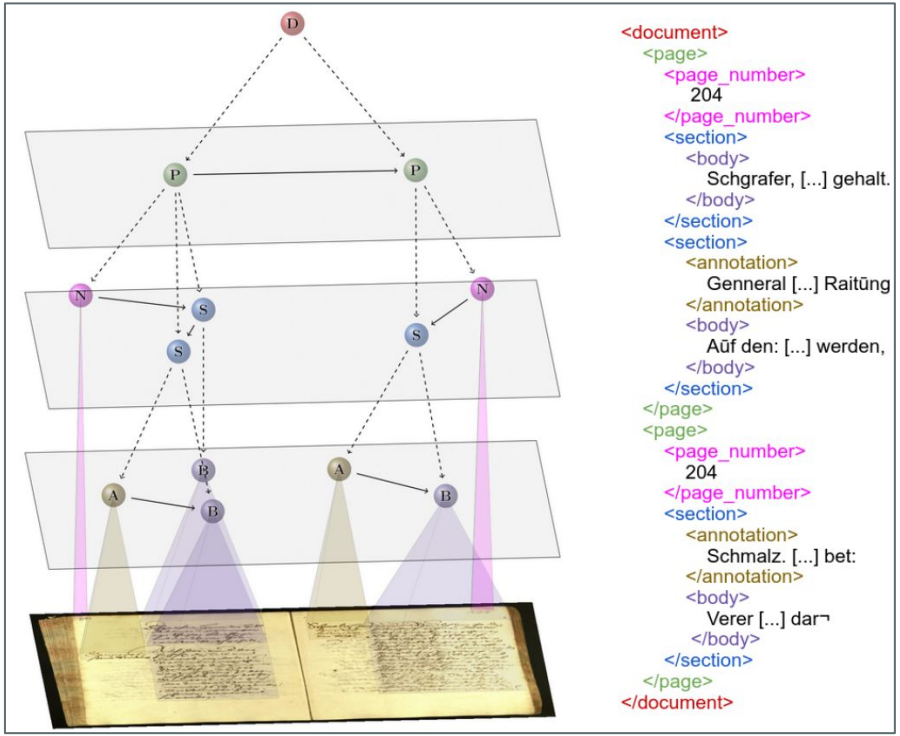
2

Handwriting recognition



2.1 Handwriting recognition with the DAN

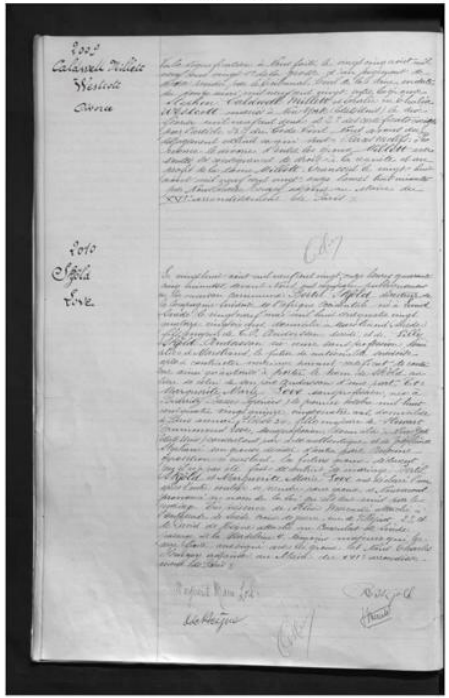
- DAN[1] (Document Attention Network)
  - handwriting recognition on a complete document of one or two pages without layout annotation
  - semantic recognition of text zones
- Objective: apply DAN to the EXO-POPP corpus to remove the text field detection step



1: D. Coquenet, C. Chatelain, T. Paquet: "DAN: a Segmentation-free Document Attention Network for Handwritten Document Recognition"

EXOPOPP_global_page_sem/valid/AD075EC_16M222_0011-left.png

Input document



0 0.2+

Attention weights

cer: 4.72% loer: 0.0% map_cer: 84.29%

Iteration: 2 **<mariage>**

Ground truth

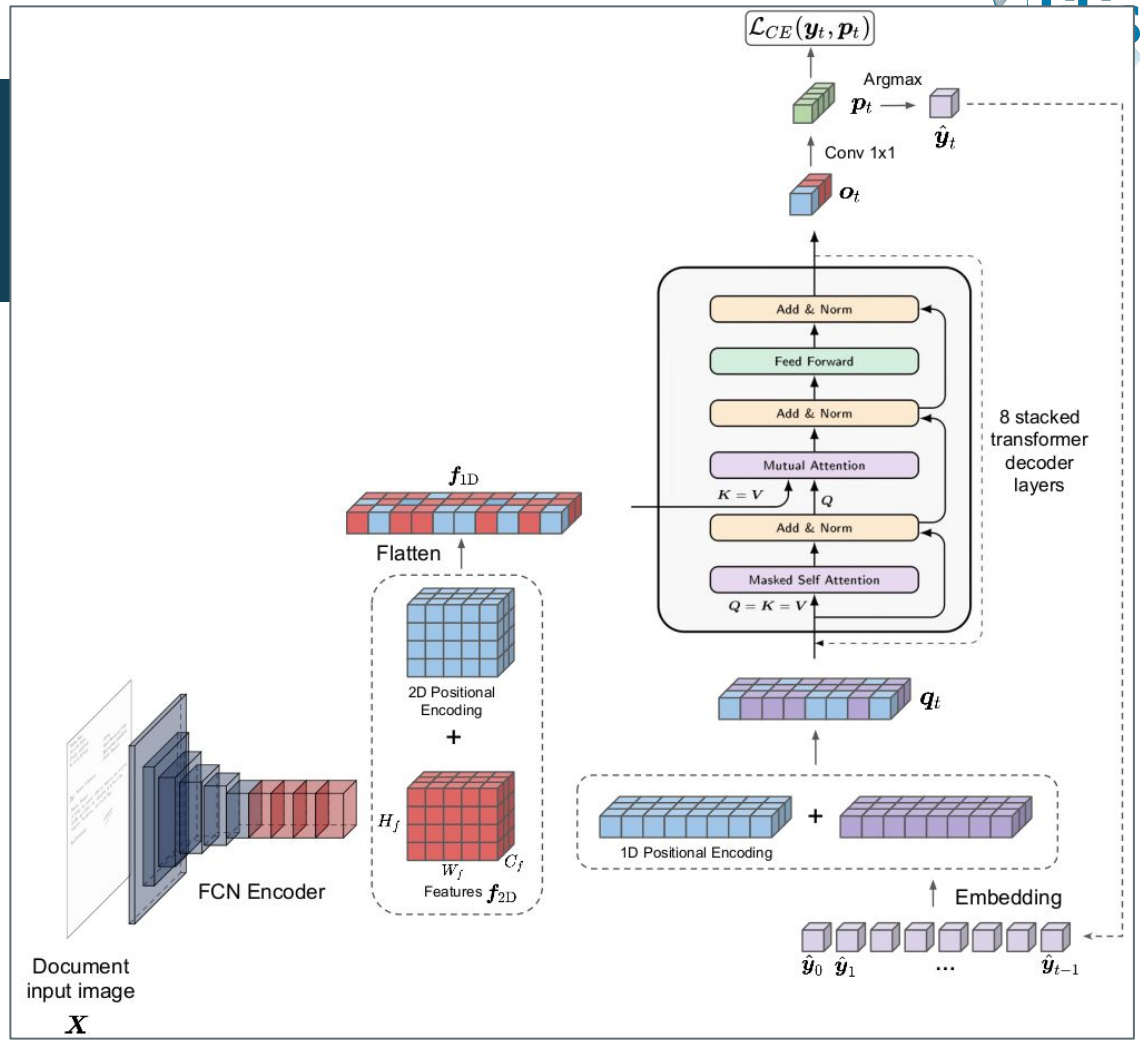
```
0
1 <page>
2   <mariage>
3     <marge-noms>
4       2009
5       Caldwell Millett
6       Westcott
7       Divorce
8     </marge-noms>
9     <paragraphe>
10      Vu la signification à Nous faite le vingt-cinq août mil
11      neuf cent vingt 1° de la grosse d'un jugement de
12      divorce rendu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

Prediction

```
0
1 <page>
2
```

2.1 Handwriting recognition with the DAN

- Encoder: same architecture as the VAN (Coquenet2022)
- 2D positional embedding of the feature map (Singh2021)
- Decoder: variant of the Transformer architecture (Vaswani2017)



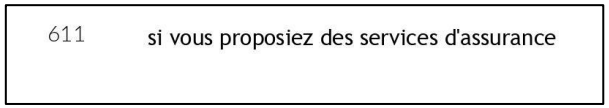
2.1 Handwriting recognition with the DAN

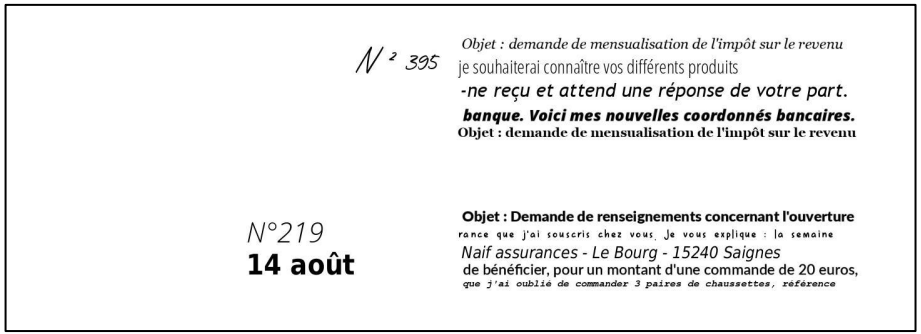
Training process:

- 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





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

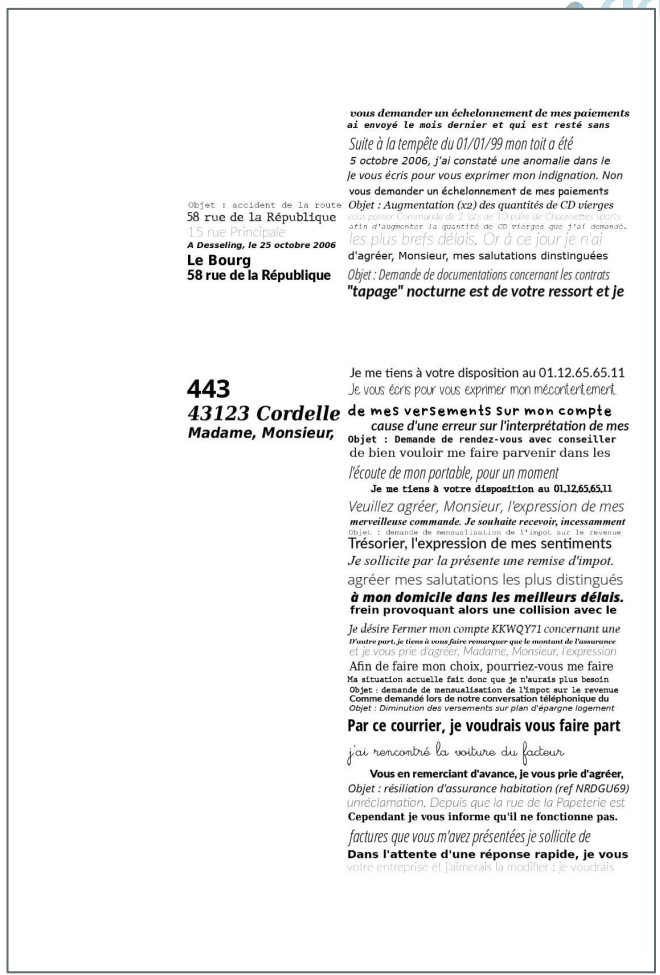
Annotated data

	Train	Valid	Test
# simple pages	254	24	24

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

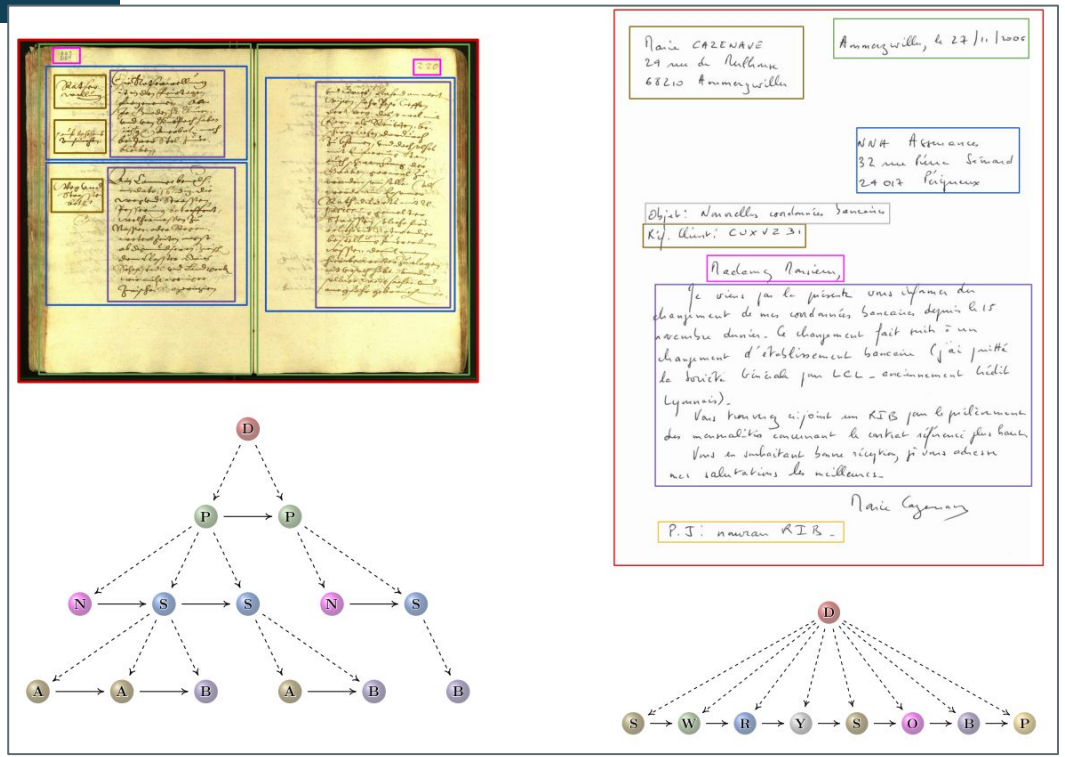


2.2 Handwriting recognition with the DAN

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

$$\mathrm{LOER} = \frac{\sum_{i=1}^{K} \mathrm{GED}(\boldsymbol{y}_i^{\mathrm{graph}}, \hat{\boldsymbol{y}}_i^{\mathrm{graph}})}{\sum_{i=1}^{K} n_{\mathrm{e}_i} + n_{\mathrm{n}_i}}$$



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.

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

Example

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

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.

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

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

2.2 Handwriting recognition Results of the DAN

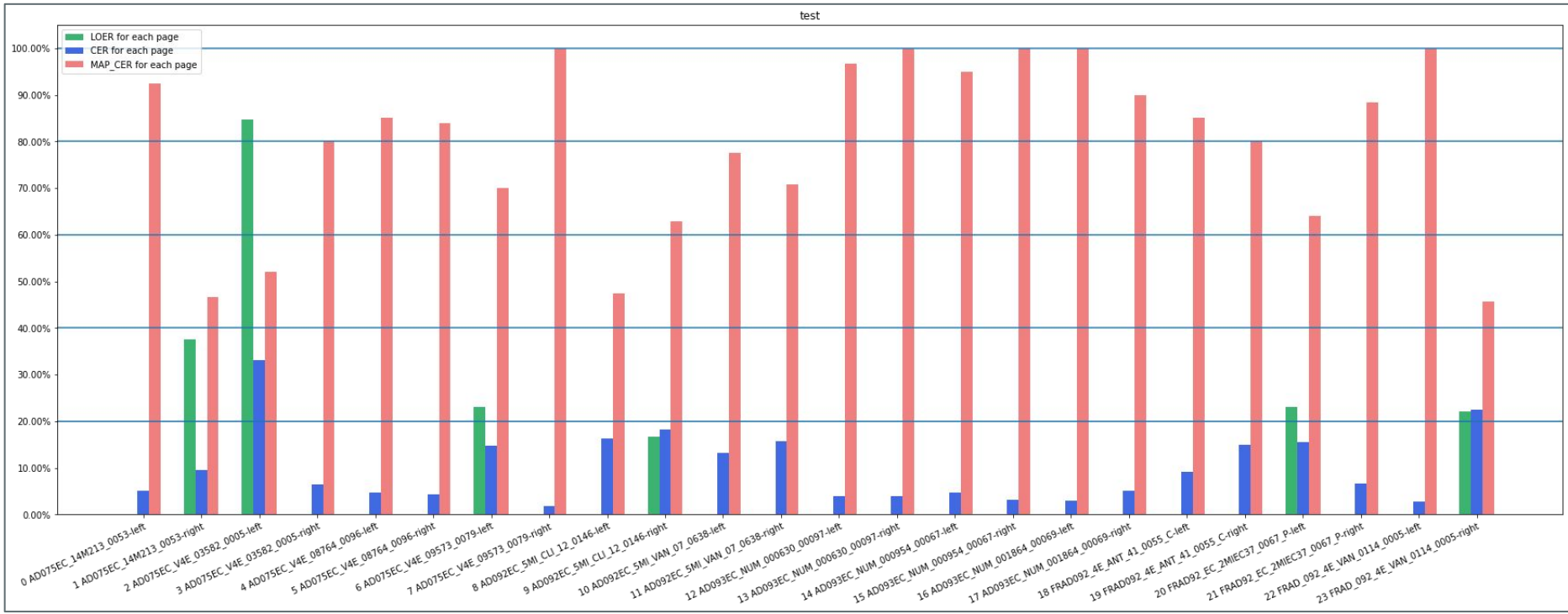
Results on Exo-POPP

	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

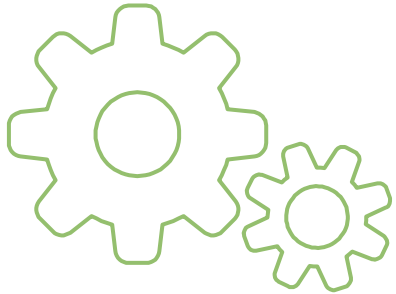
	CER	WER	LOER	MAP_CER
READ Test - simple pages	3.53%	13.33%	5.94%	92.57%
RIMES Test	4.54%	11.85%	3.82%	93.74%

2.2 Handwriting recognition Results of the DAN



3

Named entity extraction

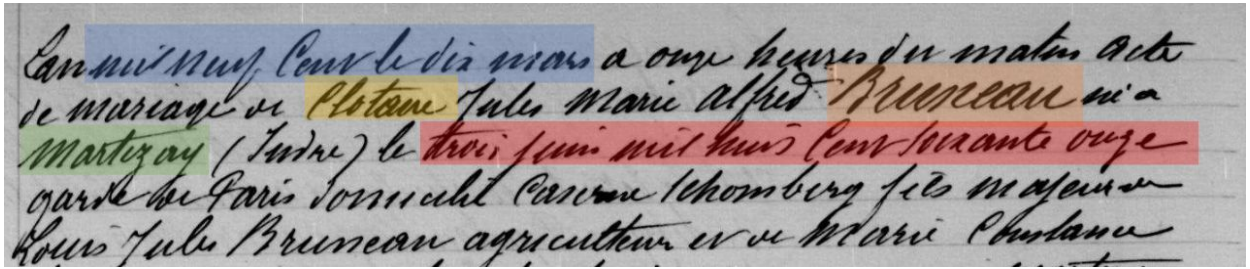


3.1 Methodology

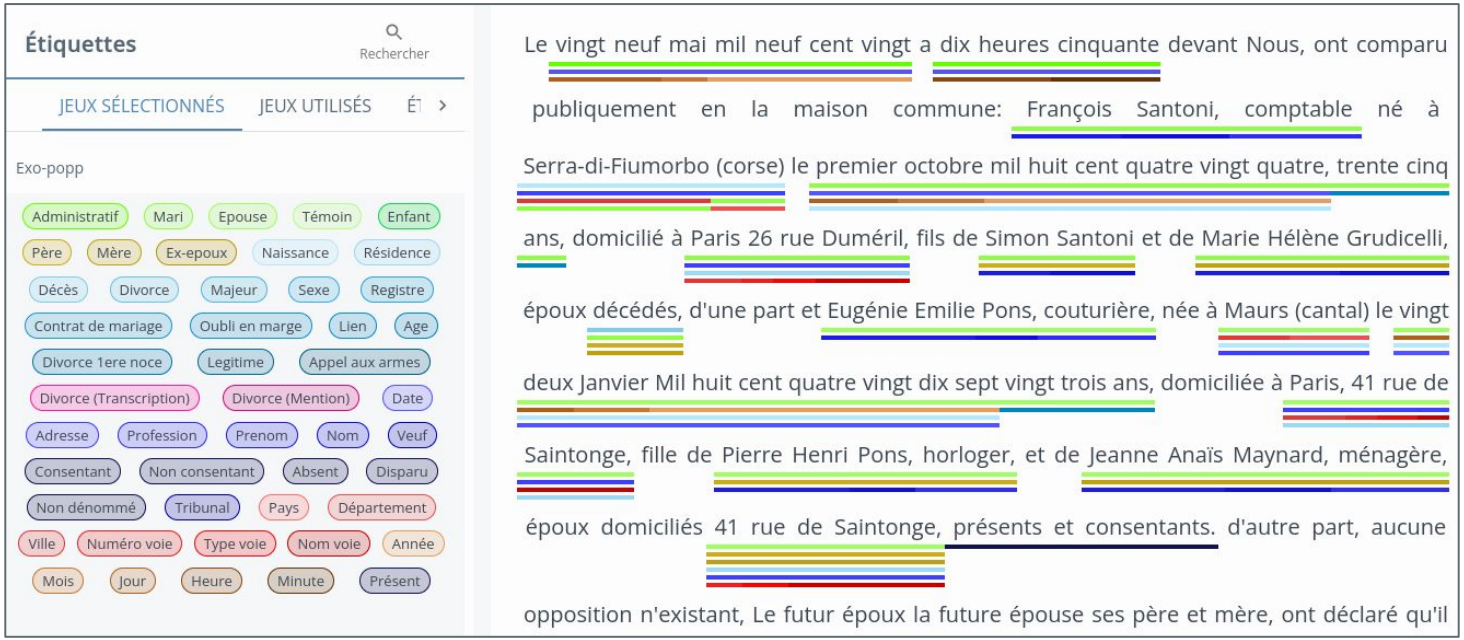
Recognition of named entities in handwritten documents

- 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".

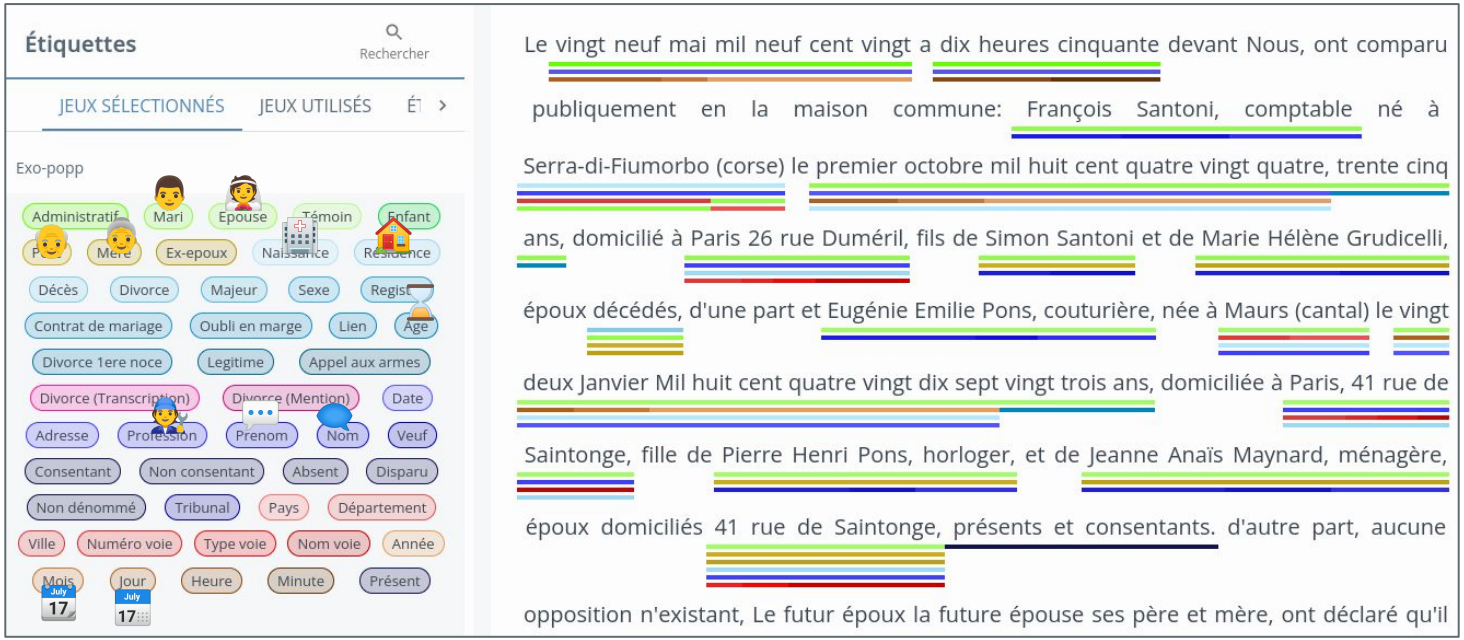
Date : 10 Mars 1900
Prénom du mari : Clotaire
Nom de famille du mari : Bruneau
Lieu de naissance du mari : Martizay
Date de naissance du mari: 3 juin mille huit cent soixante onze



3.1 Named entity recognition



3.1 Named entity recognition



3.1 Named entity recognition

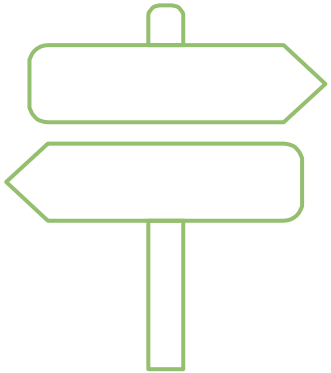
Using DAN for joint recognition of layout, handwriting and named entities

Encoding of named entities in the DAN charset:

“ⓜⓝ🗨️ 👨Santoni et 🗨️ 👰PonsⓃⓑ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ⒷⓂ”

4

Conclusion



4 Conclusion

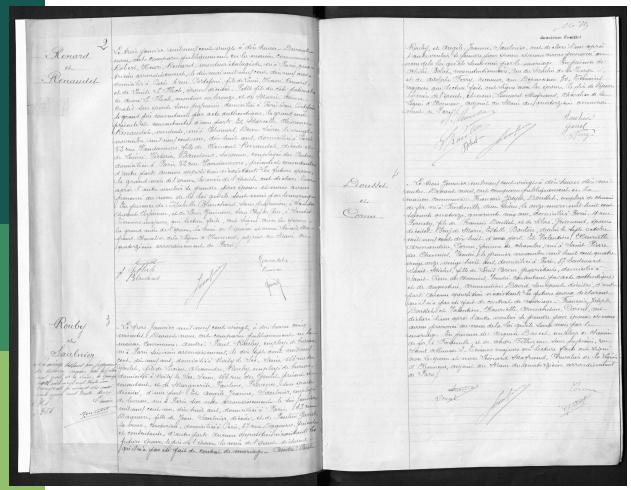
Results

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



DAN

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			...

Thank you for your attention

-

Any questions?

5.1 Bibliography

- D. Coquenet, C. Chatelain, T. Paquet: “DAN: a Segmentation-free Document Attention Network for Handwritten Document Recognition”
- D. Coquenet, 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	Nom voie
	Année	Mois	Jour	Heure	Minute	

3.3 Reconnaissance d'entités nommées

Essais d'utilisation du DAN pour la NER

Encodage des ENs dans le charset du DAN:

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