Discovering Spatial Relations in Literature: What is the Influence of OCR noise ?

Caroline Parfait (1,2,3) Gaël Lejeune (2)
Motasem Alrahabi (1,3) Glenn Roe (1,3)

Séminaires de l'équipe "Linguistique Computationnelle"

18 March 2021

caroline.parfait@sorbonne-universite.fr
gael.lejeune@sorbonne-universite.fr
(1) OBTIC, Sorbonne Université, Paris, France
(2) STIH, Sorbonne Université, Paris France
(3) SCAI, Sorbonne Center for Artificial Intelligence, Paris, France

Table of Contents

Quel(s) but(s) ?

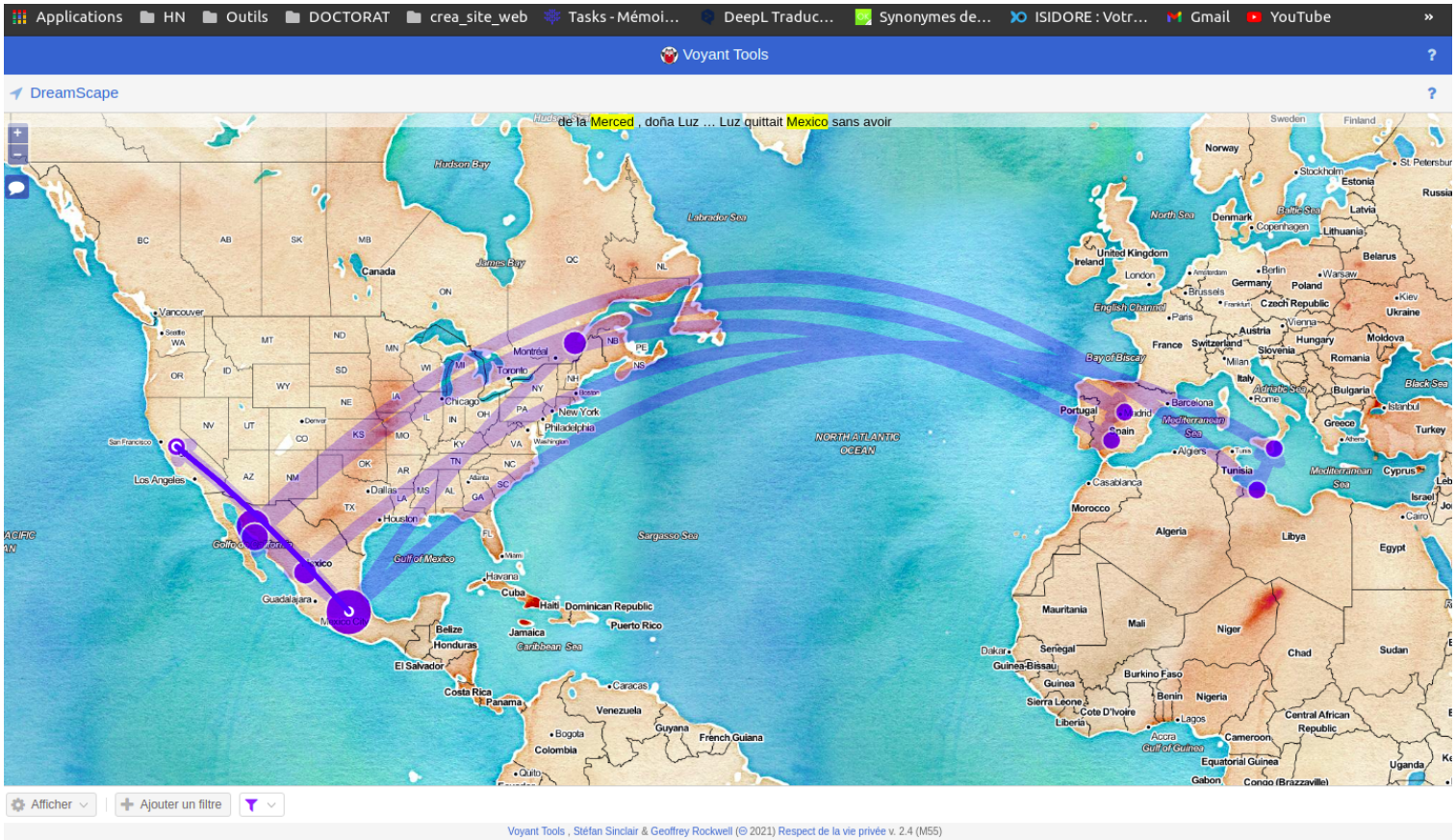


Figure: Voyant Tools, *"Les trappeurs de l'Arkansas (4e édition) "*,Gustave Aimard, 1858.

Introduction

Phd Subject

Literary space analysis, three dimensions

Data
- Literature VS News
- OCR data quality and its influence

Introduction

Phd Subject

Literary space analysis, three dimensions

Data	• Literature VS News • OCR data quality and its influence
Task/users	• What do the users want ? • Do we evaluate it properly with P/R/F-score ?

Introduction

Phd Subject

Literary space analysis, three dimensions

Data	• Literature VS News • OCR data quality and its influence
Task/users	• What do the users want ? • Do we evaluate it properly with P/R/F-score ?
Methods	• Adaptability of Named-Entity Recognition systems • Complementarity between these systems

Introduction

Phd Subject

Literary space analysis, three dimensions

Data	• Literature VS News • OCR data quality and its influence
Task/users	• What do the users want ? • Do we evaluate it properly with P/R/F-score ?
Methods	• Adaptability of Named-Entity Recognition systems • Complementarity between these systems

- How to overcome the limitations of NER quality ?
- Do NER tools meet the expectation of users ?

Introduction

Proposed approach : "user-centred design"

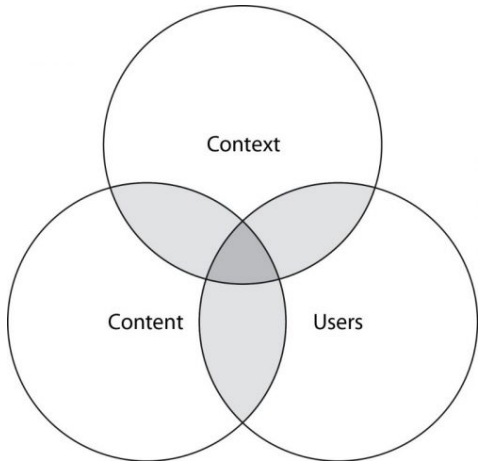


Figure: Peter Morville, *"Three Circles of Information Architecture"*, 2004

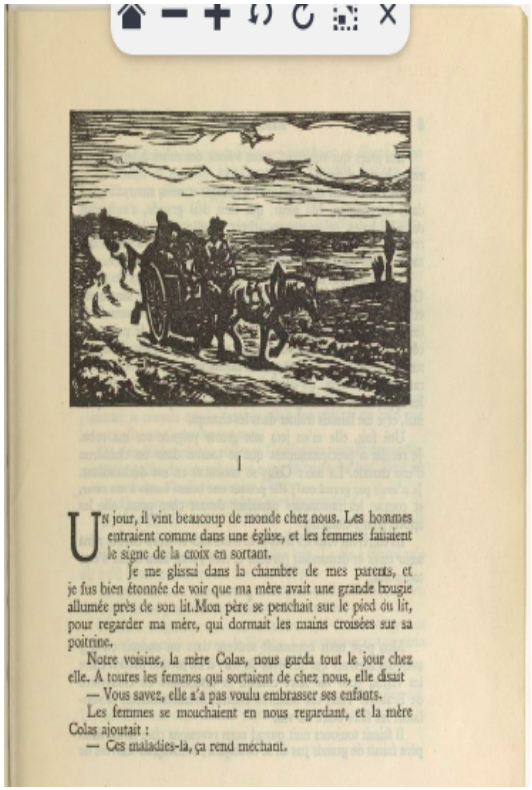
- **Task** : Spatial NER in **imperfect conditions**
- **Users** : Researchers in **various fields**
- **Data** : A rather heterogeneous corpus of French novels (19th/20th cent.) $\rightarrow$ **variability** ?

Introduction

Corpus

Références	**nb mots**	**nb EN PP**	**nb EN OCR**
- *"Je suis un aventurier"*, de Jacques Dutronc, 1970.	213	24	23
- *"Dictionnaire géographique"*, Girault de Saint-Fargeau, Volume 1, 1844-1846.	1400	160	165
- *"Marie-Claire"*,Marguerite Audoux, 1925.	34 800	180	300
- *"Les trappeurs de l'Arkansas (4e édition) "*,Gustave Aimard, 1858.	87542	1008	1335

Correlation between input quality and output quality ?



I

Un jour, il vint beaucoup de monde chez nous. Les hommes entraient comme dans une église, et les femmes faisaient le signe de la croix en sortant.

Je me glissai dans la chambre de mes parents, et je fus bien étonnée de voir que ma mère avait une grande bougie allumée près de son lit. Mon père se penchait sur le pied du lit, pour regarder ma mère, qui dormait les mains croisées sur sa poitrine.

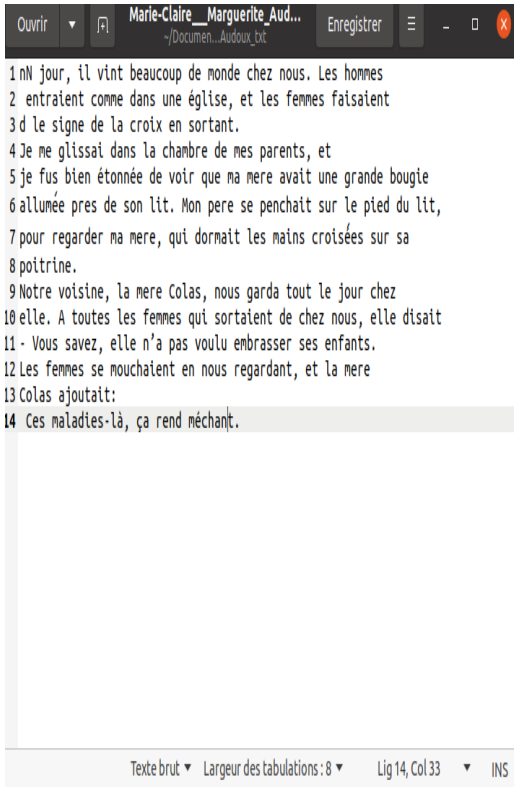
Notre voisine, la mère Colas, nous garda tout le jour chez elle. A toutes les femmes qui sortaient de chez nous, elle disait

— Vous savez, elle n'a pas voulu embrasser ses enfants.

Les femmes se mouchaient en nous regardant, et la mère Colas ajoutait :

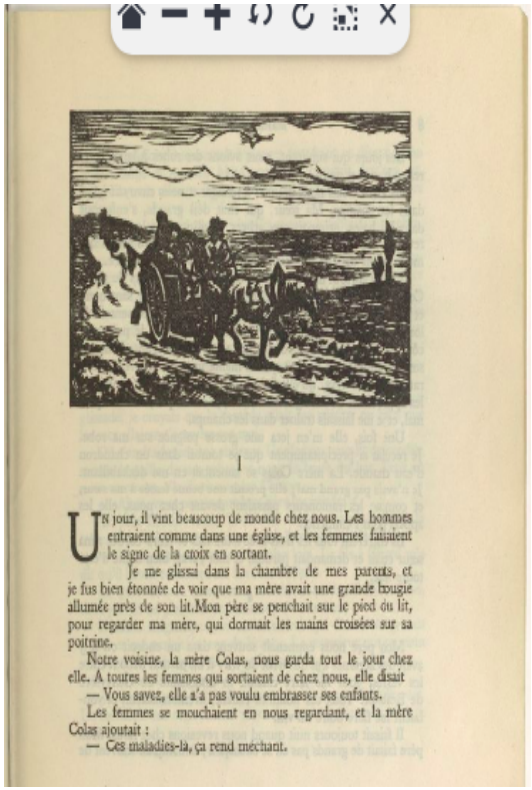
— Ces maladies-là, ça rend méchant.

(a) Image from Gallica



(b) Text after OCR processing

Correlation between input quality and output quality ?



I

Un jour, il vint beaucoup de monde chez nous. Les hommes entraient comme dans une église, et les femmes faisaient le signe de la croix en sortant.

Je me glissai dans la chambre de mes parents, et je fus bien étonnée de voir que ma mère avait une grande bougie allumée près de son lit. Mon père se penchait sur le pied du lit, pour regarder ma mère, qui dormait les mains croisées sur sa poitrine.

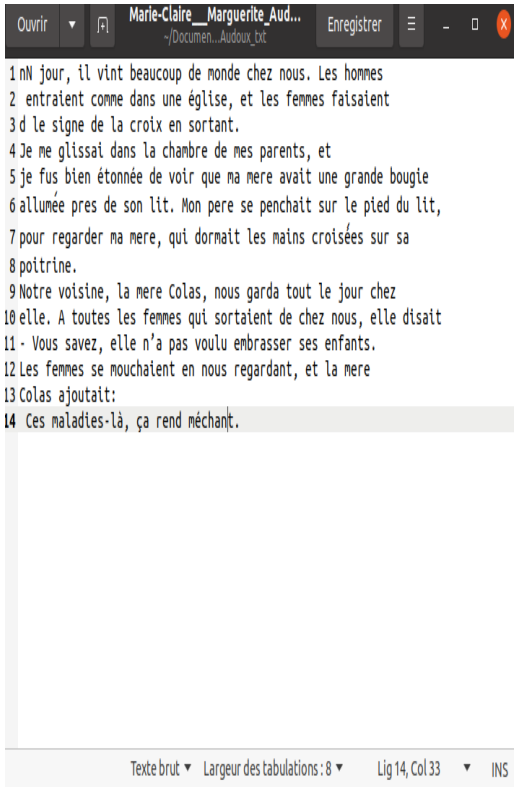
Notre voisine, la mère Colas, nous garda tout le jour chez elle. A toutes les femmes qui sortaient de chez nous, elle disait

— Vous savez, elle n'a pas voulu embrasser ses enfants.

Les femmes se mouchaient en nous regardant, et la mère Colas ajoutait :

— Ces maladies-là, ça rend méchant.

(a) Image from Gallica



(b) Text after OCR processing

- Distance in inputs (ELTeC version VS OCR)
- Distance in outputs (NER on ELTeC version VS NER on OCR data)?
- Causality between noise in input and noise in output ?

Explicit comparisons for humans

Data : "Je suis un aventurier", J.Dutronc (NER with Spacy SM)

- Exp. 1 : NER output for two texts with many NE
- Exp. 2 : Comparisons of distances between input and the NER output

Explicit comparisons for humans

Data : "Je suis un aventurier", J.Dutronc (NER with Spacy SM)

- Exp. 1 : NER output for two texts with many NE
- Exp. 2 : Comparisons of distances between input and the NER output

Clean Text	**NER**	**Noisy text**	**NER**
J'ai fait la vie à Varsovie	Varsovie	J'ai fait la vie Varsovie.	Varsovie (TP)
J'ai fait le rat à Canberra	Canberra	J'ai fait le rat a Camberra.	Camberra (FP?)
J'ai fait des games à Birmingham	Birmingham	J'ai fait des games à Binningham.	'**PER**' (FP)

Explicit comparisons for humans

"Je suis un aventurier", J.Dutronc, Spacy SM

"Je suis un aventurier", J.Dutronc, Spacy SM					
Clean Text	**ENG**	**Good OCR**	**ENG**	**More noise**	**ENG**
J'ai fait le soldat à Bogota	(Bogota,)	J'ai fait l'soldat à Bogola	(Bogola,)	J'ai fait l'soldat à Bogola	(Bogola,)
J'ai eu la berlue à Berlin	(Berlin,)	J'ai eu la ber1ue à berlin.	(ber1ue,)	J'ai eu la ber1ue à berlin.	(ber1ue,)
J'ai fait la vie à Varsovie	(Varsovie,)	J'ai fait la vie à Varsovie	(Varsovie,)	J'ai fait la vie à varsovie	()

Table

Explicit comparisons for humans

"Dictionnaire géographique", Girault de Saint-Fargeau (Vol. 1)

n°	**Text**	**EN-S**	**OCR Text**	**EN-S**
1	ÉVERLY, vg. Seine-et-Marne (Brie), arr.	(Seine-et-Marne, Brie)	EVERLY, vg. (Brie) , arr.	**(EVERLY, Brie)**
2	ÉVEUX, vg. Rhône (Lyonnais), arr. et à	(Rhône, Lyonnais)	ÉVEUX, vg. Rhr;ne (Lyonnais), arr. et å	**(Rhr;ne**, Lyonnais)
...	...	...	...	

Table

Explicit comparisons for humans

"Dictionnaire géographique", Girault de Saint-Fargeau (Vol. 1)

n°	**Text**	**EN-S**	**OCR Text**	**EN-S**
3	ÉVILLERS, vg. Doubs (Franche-Comté),	(ÉVILLERS, Doubs, Franche, Comté)	ÉVILLF.RS, vg. Doubs (Franche-Comté),	**(ÉVILLF.RS, Doubs, Franche, Comté)**
4	ÉVIN-MALMAISON, vg. Pas-de-Calais	(ÉVIN, MALMAI-SON, vg, Pas-de-Calais)	, vg, Pas—de-Calais	**(vg, Pas—de-Calais)**
...	...	...	...	

Explicit comparisons for humans

"Les trappeurs de l'Arkansas (4e édition) ",Gustave Aimard, 1858 .

Clean Text	**NER**	**OCR Text**	**NER**
Le voyageur qui pour la première fois débarque dans l'Amérique du Sud éprouve malgré lui un sentiment de tristesse indéfinissable	(Amérique du Sud, 'LOC')	Le voyageur qui pour la premiere fois **débarquedans** l'Amerique du Sud, éprouve malgré lui unsentiment de tristesse indefinissable	(du Sud, 'LOC')
...	...	...	...

Explicit comparisons for humans

"Les trappeurs de l'Arkansas ",Gustave Aimard, 1858 .

Clean Text	**NER**	**OCR Text**	**NER**
C'est à ce déplacement continuel qu'il faut attribuer, en Amérique, l'absence de ces grands monuments,	Amérique	C'est à ce déplacement continuel qu'il faut attri buer, en Amérique, l'absence de ces grands monu ments,	Amérique
... et ces milliers d'animaux que la civilisation des autre parties de **l'Amérique** refoule de jour en jour...	**()**	... et ces milliersd'animaux que la civilisation des autre parties **del'Amerique** refoule de jour en jour...	**del'Amerique**
...	...	...	...

Explicit comparisons for humans

Les trappeurs de l'Arkansas, Gustave Aimard, 1858 .

Clean Text	**NER**	**OCR Text**	**NER**
Ce vaste continent, qui pendant trois siècles a été la paisible possession des Espagnols, [...]avant la découverte, par les Aztèques et les Incas sont encore debout dans leur majestueuse simplicité	(Espagnols, Incas)	Ce vaste continent, qui pendant trois siecles a étéla paisible possession des Espagnols, [...] avant **ladécouverte**, par les **Azteques** et les **Ineas** sont encoredebout dans leur majestueuse simplicit,	(Espagnols, **Azteques**, **Ineas**)
...	...	...	...

Table

Explicit comparisons for humans

Some results

Data

- Despite spelling mistakes the tool finds a spatial NE
- System uses Syntactic information, not only the lexicon

Explicit comparisons for humans

Some results

Data	• Despite spelling mistakes the tool finds a spatial NE • System uses Syntactic information, not only the lexicon
Task/users	• Can the user be pleased/confident with the output? • An out-of-the-box NER tools.

Explicit comparisons for humans

Some results

Data	• Despite spelling mistakes the tool finds a spatial NE • System uses Syntactic information, not only the lexicon
Task/users	• Can the user be pleased/confident with the output? • An out-of-the-box NER tools.
Method	• How can the tool help to improve the user's feeling ?

First test of automatic analysis - Precision/Recall/F-score

"Je suis un aventurier", de Jacques Dutronc, 1970

Tableau

Entrée [5]:
```
tab = pd.DataFrame({"Precision":[precision_liste[0],precision_liste[1],precision_liste[2
tab
```

Out[5]:

	Precision	Recall	F-Mesure
fr_core_news_sm	95.0	86.363636	90.476190
fr_core_news_lg	95.0	90.476190	92.682927
fr_core_news_md	95.0	86.363636	90.476190

Distances and similarities: Automatic comparisons

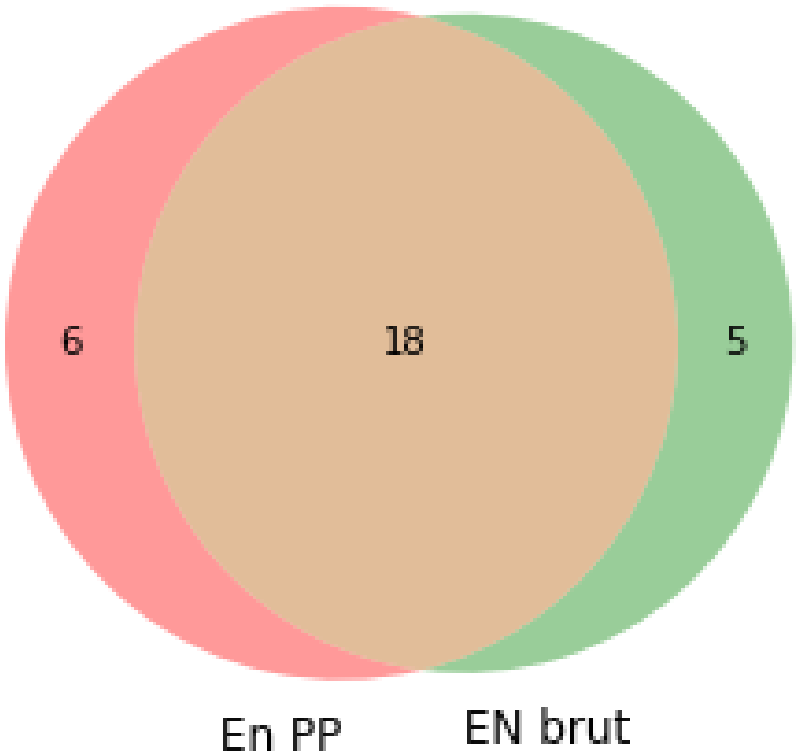
M. AUDOUX, nb mots = 34 8000, nb EN = 300 OCR/180 PP			
Modèle de langue testé	**Precision**	**Recall**	**F-Mesure**
fr_core_news_sm	9.375000	100.0	17.142857
fr_core_news_lg	32.500000	100.0	49.056604
fr_core_news_md	26.666667	100.0	42.105263
...	...	...	...

Table: Distance/Similarity between outputs (NER on clean data VS NET on OCR data)

First test of automatic analysis

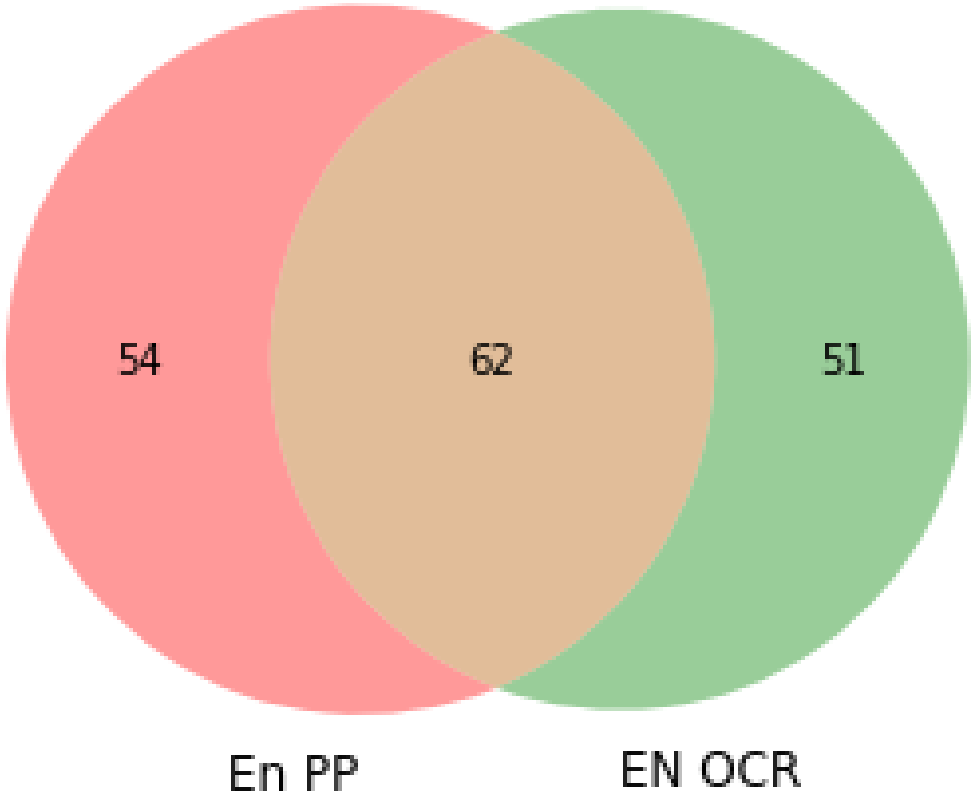
"Je suis un aventurier", de Jacques Dutronc, 1970

Intersection des EN détectées dans le texte propre et dans le texte ocr, Dutronc



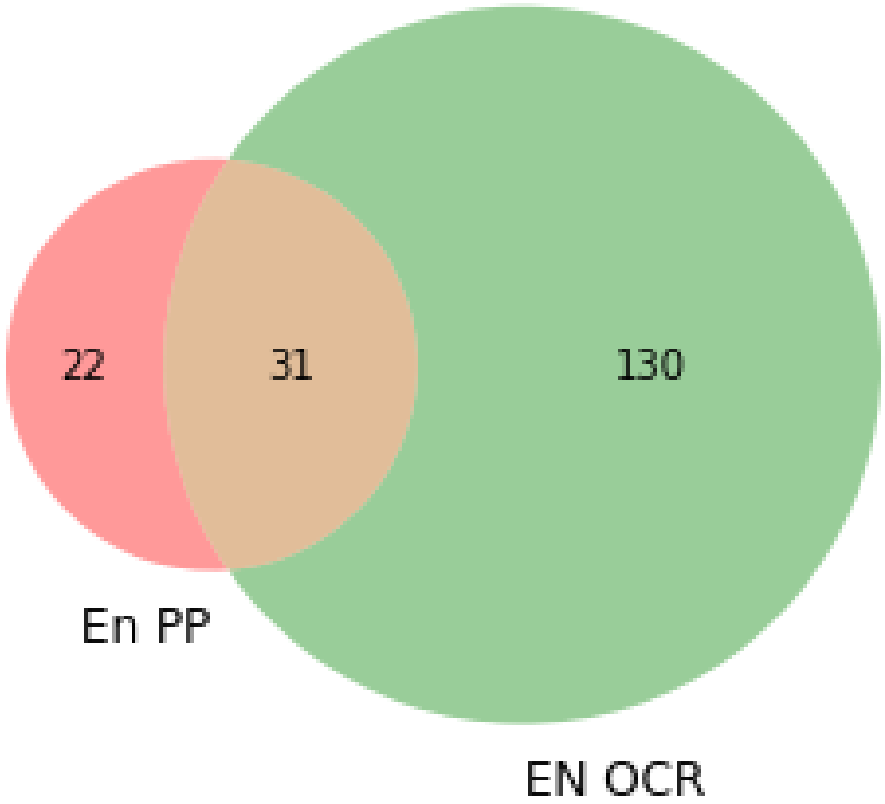
First test of automatic analysis

Intersection des EN détectées dans le texte propre et dans le texte ocr, Dico Gé



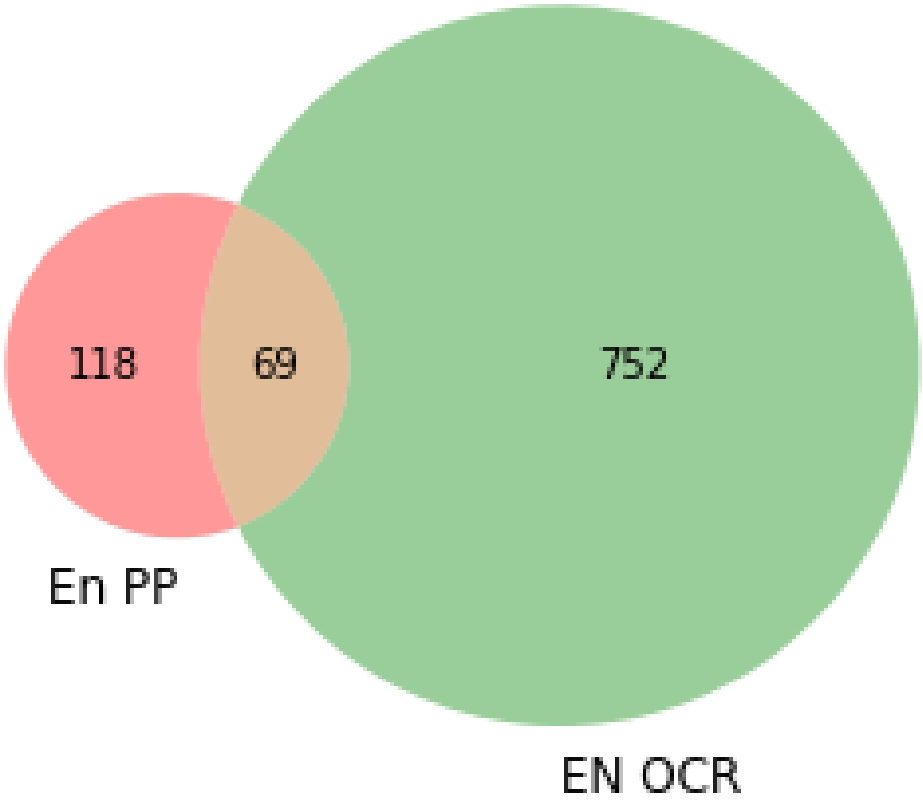
First test of automatic analysis

Intersection des EN détectées dans le texte propre et dans le texte ocr, Audoux



First test of automatic analysis

Intersection des EN détectées dans le texte propre et dans le texte ocr, Aimard



Distances and similarities: Automatic comparisons

Distances

- $Jaccard = \frac{|set(V) \cap set(W)|}{|set(V) \cup set(W)|}$
- $Dice = \frac{2 * |set(V) \cap set(W)|}{|set(V) \cup set(W)|}$
- $Cosinus = \frac{V.W}{||V||.||W||}$
- $Bray - Curtis = \frac{2 \sum_{i=1}^{n} min(V[i], W[i])}{\sum_{i=1}^{n} (V[i] + W[i])}$ avec $vw = min(V_i, W_i)$ pour chaque dimension i

Distances and similarities: Automatic comparisons

"Je suis un aventurier" Jacques Dutronc, 1970

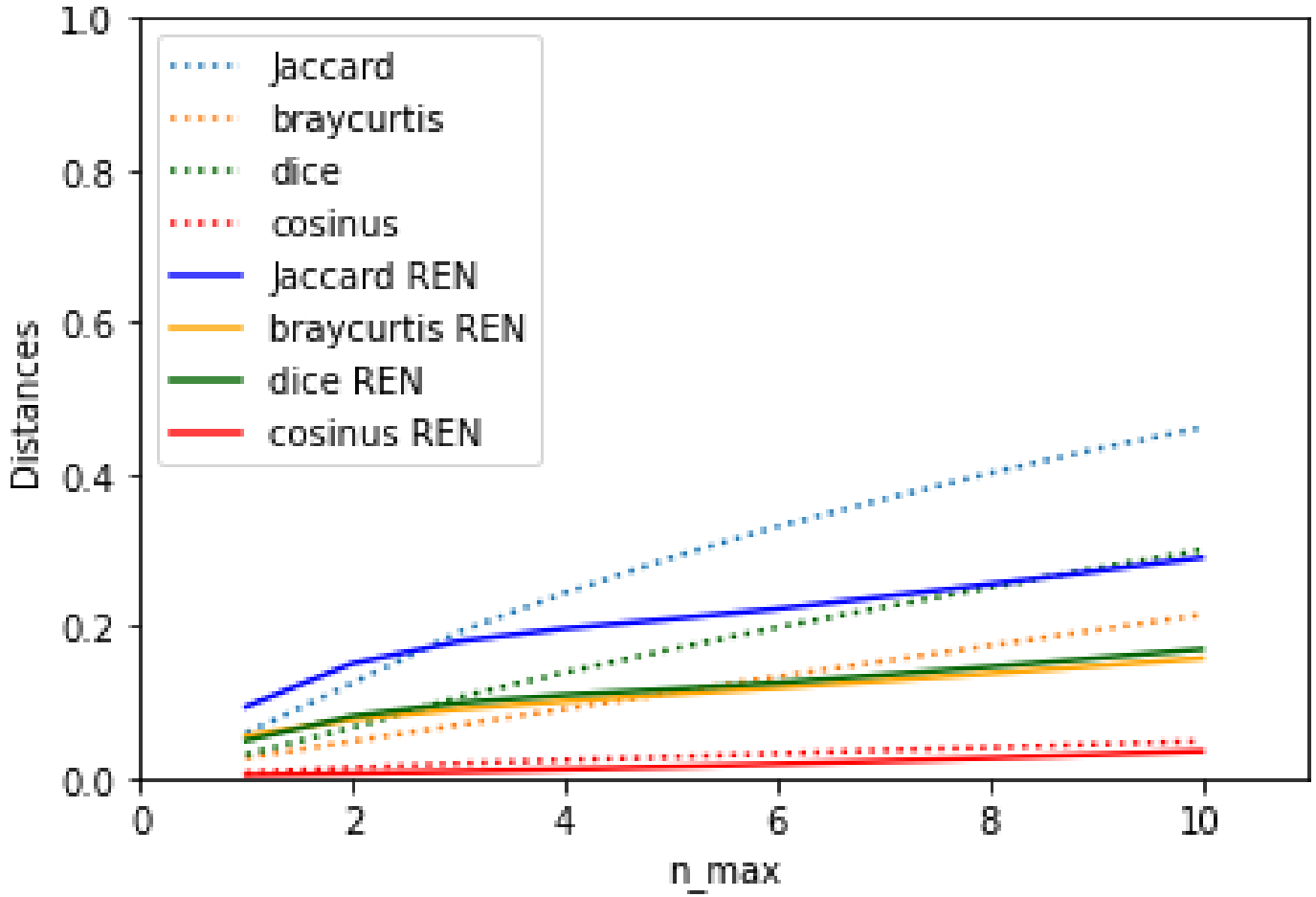


Figure: Distances in the input ("Official text" VS OCR **WER = 0.0.234**) and in the output (NER on clean data VS NER on OCR **WER=0.25**)

Distances and similarities: Automatic comparisons

"Je suis un aventurier" Jacques Dutronc, 1970

Distance[1]	**Dist "word"**	**sim**	**Dist "char, nmax=5"**	**sim**
Jaccard	0.3548		0.2090	
Dice	0.2156	N/A	0.1166	N/A
Bray curtis	0.2142	0.7917	0.1098	
Cosinus	0.1818		0.0144	
wer	0.25	N/A	N/A	N/A
cer	0.1340	N/A	N/A	N/A
...	...	...	...	...

Table: Distance/Similarity between the NER output (Official Text VS Noisy Text), NER model SPACY_SM

[1] Obtained with SKLEARN

Distances and similarities: Automatic comparisons

"Je suis un aventurier" Jacques Dutronc, 1970

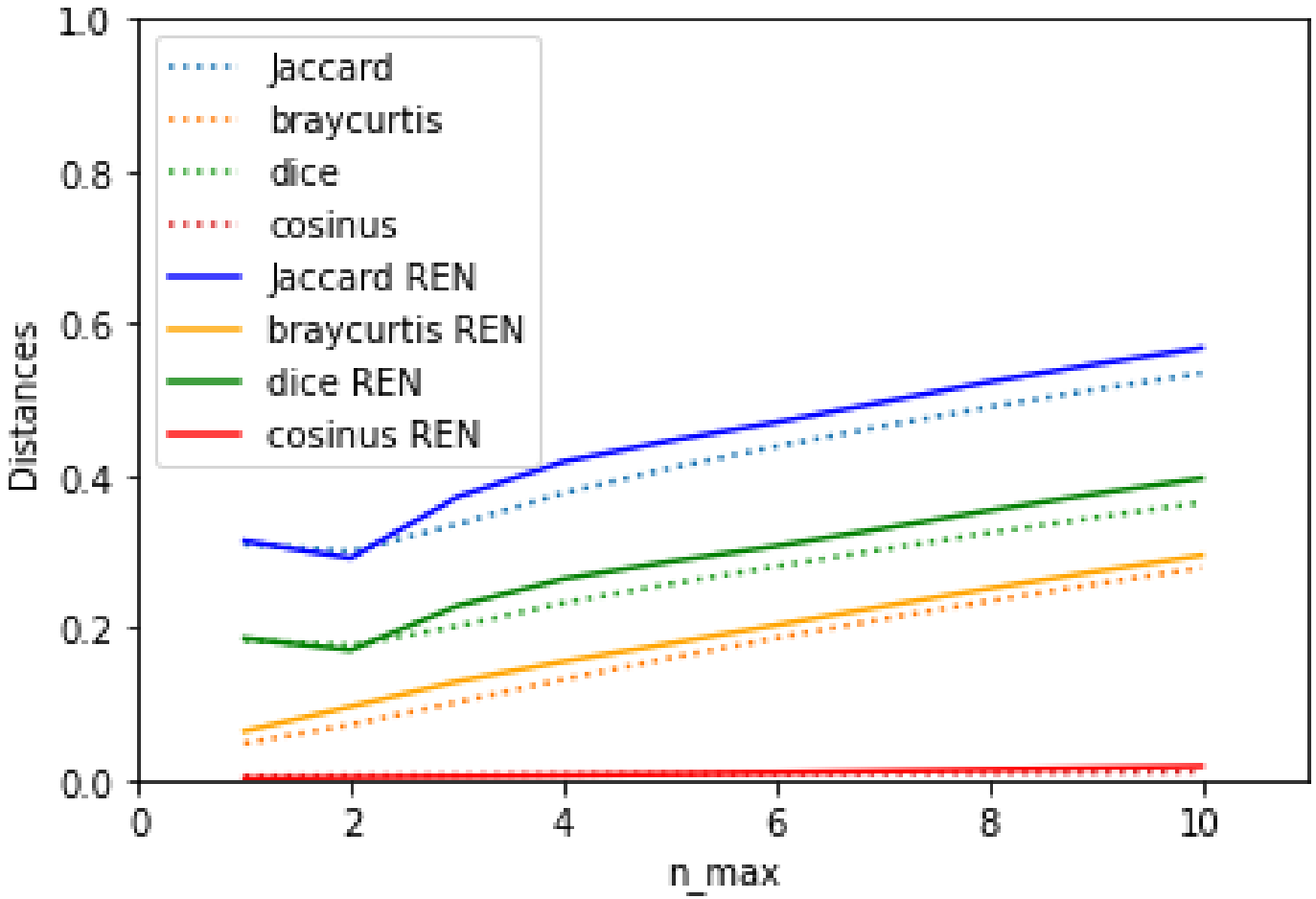
Distance[2]	**Dist "word"**	**sim**	**Dist "char, nmax=5"**	**sim**
Jaccard	0.1363	0.8637	0.2895	
Dice	0.07317	N/A	0.1693	N/A
Bray curtis	0.0481	0.9519	0.1119	
Cosinus	0.0106	N/A	0.0274	N/A
Wer	0.2347	N/A	N/A	N/A
Cer	0.0642	N/A	N/A	N/A
...	...	...	...	...

Table: Distance/Similarity between Official Text and Noisy Text

[2] sklearn

Distances and similarities: Automatic comparisons

"Dictionnaire géographique", Girault de Saint-Fargeau, (Vol. 1)



Distances and similarities: Automatic comparisons

"Dictionnaire géographique", Girault de Saint-Fargeau (Vol. 1)

Distance[3]	**Dist "word"**	**sim**	**Dist "char, nmax=5"**	**sim**
Jaccard	0.5389	0,4611	0.4456	
Dice	0.3688	N/A	0.2867	N/A
Bray curtis	0.3588	0.6469	0.1790	
Cosinus	0.1868		0.0076	
Wer	0.5493	N/A		
Cer	0.3364	N/A		
...	...	...	...	...

Table: Distance/Similarity in the input (Gallica Text VS OCR Text) and in the output (NER on each data)

[3] sklearn

Distances and similarities: Automatic comparisons

"Dictionnaire géographique", Girault de Saint-Fargeau (Vol. 1)

Distance[4]	**Dist "word"**	**sim**	**Dist "char, nmax=5"**	**sim**
Jaccard	0.4031	0.5968	0.4093	
Dice	0.2524	N/A	0.2573	N/A
Bray curtis	0.1789	0.8211	0.1599	
Cosinus	0.0173		0.0071	
Wer	0.2652	N/A	N/A	N/A
Cer	0.0963	N/A	N/A	N/A
...	...	...	...	...

Table: Distance/Similarity between Gallica Text and OCR Text

[4] sklearn

Distances and similarities: Automatic comparisons

"Marie-Claire",Marguerite Audoux, 1925.

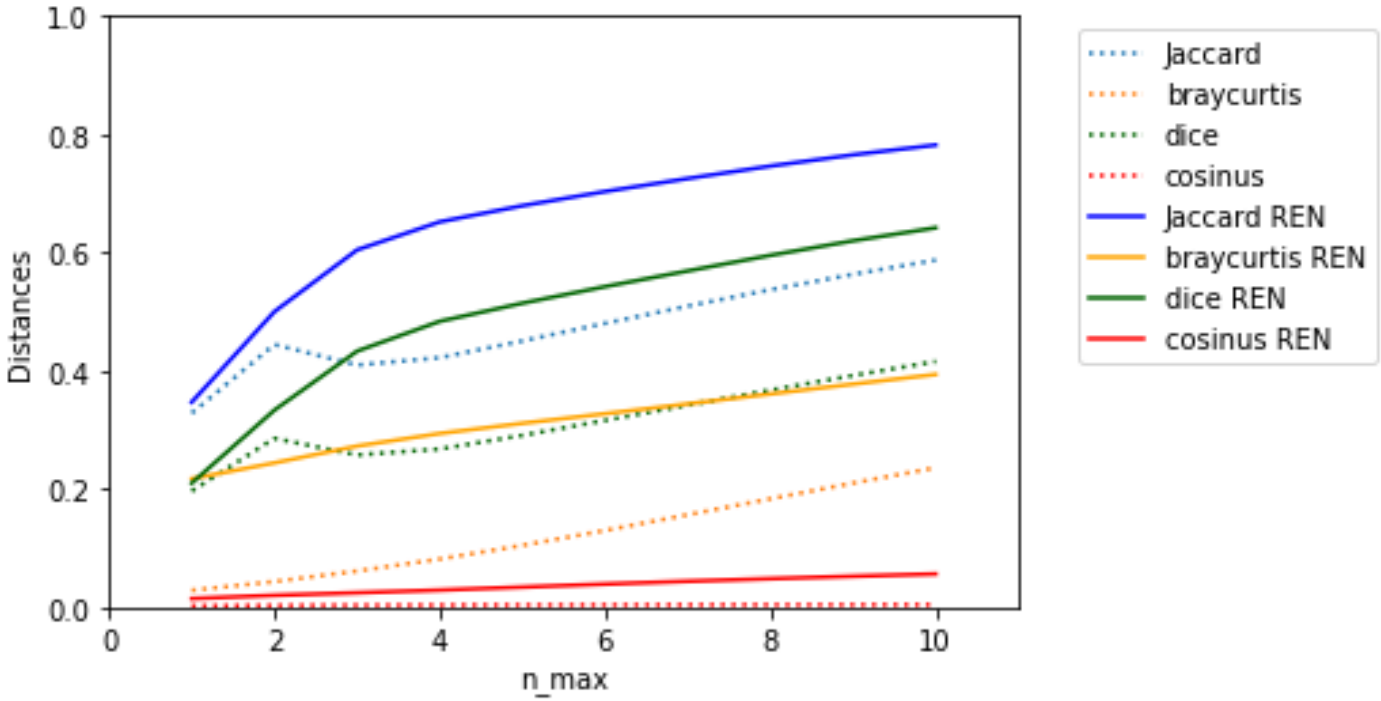


Figure: Distances in input (ELTEC VS OCR) and NER output

Distances and similarities: Automatic comparisons

"Marie-Claire",Marguerite Audoux, 1925.

Distances between NER output in ELTeC texts and OCR texts				
Distance[5]	**Dist "word"**	**sim**	**Dist "char, nmax=5"**	**sim**
Jaccard	0.7777	0.2223	0.6792	0.3208
Dice	0.6363	N/A	0.5142	N/A
Bray curtis	0.4442		0.3110	0.6890
Cosinus	0.1383	0	0.0339	
Wer	1.1683	N/A		
Cer	0.7818	N/A		
...	...	...	...	...

Table: Distance/Similarity between ELTeC Text and OCR Text ; NER Tool and Model : Spacy_sm

[5] sklearn

Distances and similarities: Automatic comparisons

"Marie-Claire",Marguerite Audoux, 1925.

Distances between ELTeC texts and OCR texts				
Distance[6]	**Dist "word"**	**sim**	**Dist "char, nmax=5"**	**sim**
Jaccard	0.3144		0.4504	
Dice	0.1865	N/A	0.2906	N/A
Bray curtis	0.0829		0.1045	
Cosinus	0.0043		0.0036	
Wer (10% du texte)	0.1992	N/A		
Cer	¿¿	N/A		
...	...	...	...	...

Table: Distance/Similarity between ELTeC Text and OCR Text ; OCR Tool : Kraken

[6] L'Heure

Distances and similarities: Automatic comparisons

Les trappeurs de l'Arkansas, Gustave Aimard, 1858.

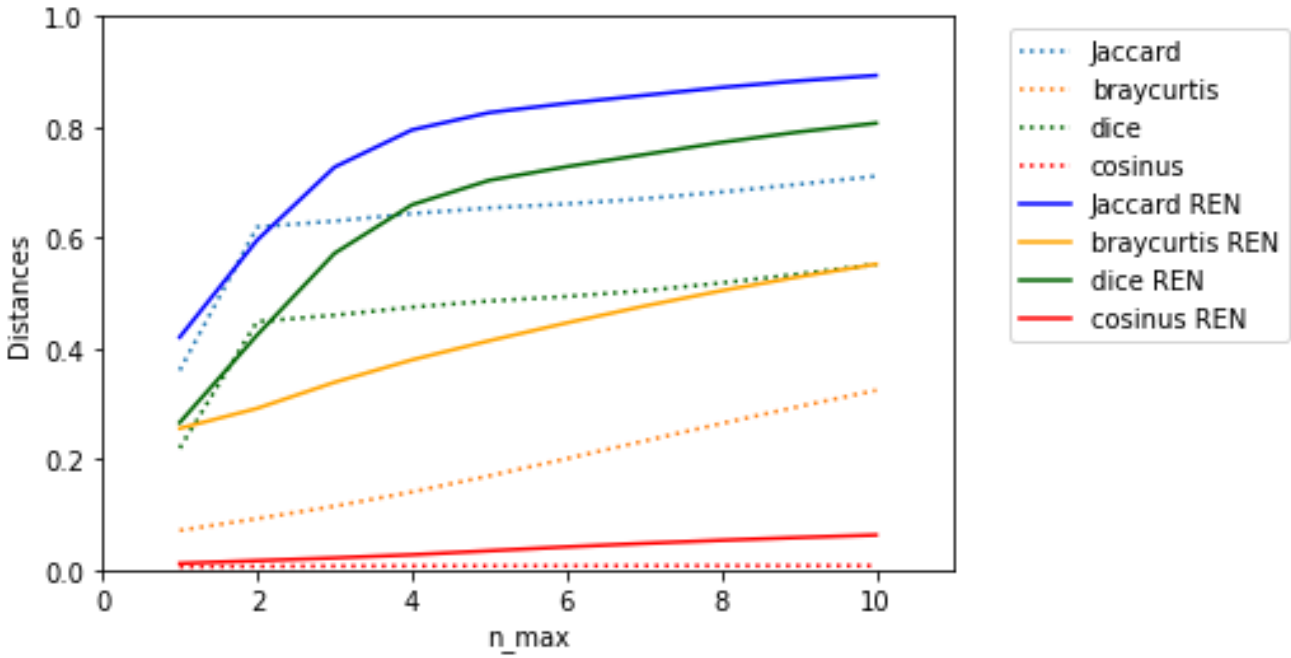


Figure: Distances in input (ELTEC VS OCR) and NER output

Distances and similarities: Automatic comparisons

Les trappeurs de l'Arkansas, Gustave Aimard, 1858.

Distance[7]	**Dist "word"**	**sim**	**Dist "char, nmax=5"**	**sim**
Jaccard	0.9130	0,0870	0.8260	0,1740
Dice	0.8400	N/A	0.7036	N/A
Braycurtis	0.6715	0,3285	0.4084	0,5916
Cosinus	0.3286	0	0.0343	
Wer	1.2248	N/A	N/A	N/A
Cer(1 EN/I)	0.8966	N/A	N/A	N/A
...	...	...	...	...

Table: Distance/Similarity between outputs (NER on clean data VS NET on OCR data)

[7] sklearn

Distances and similarities: Automatic comparisons

"Les trappeurs de l'Arkansas (4e édition) ",Gustave Aimard, 1858.

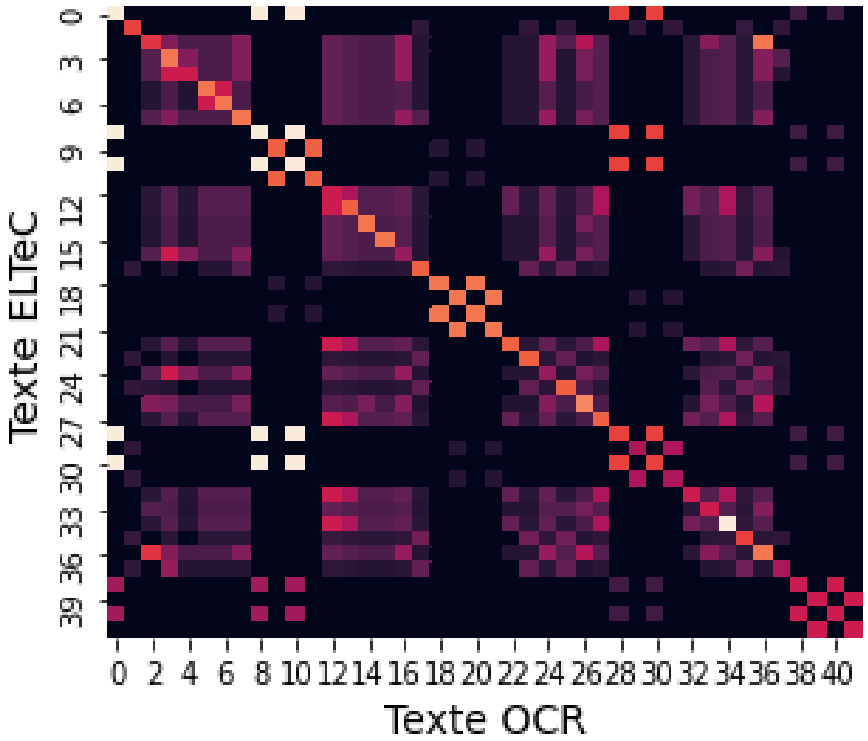
Distance[8]	**Dist "word"**	**sim**	**Dist "char, nmax=5"**	**sim**
Jaccard	0.6942		0.6538	
Dice	0.5316	N/A	0.4856	N/A
Bray curtis	0.2357		0.1697	
Cosinus	0.0097	0	0.0071	
Wer (10% du texte)	0.3836	N/A		
Cer (10% du texte)	¿¿	N/A		
...	...	...	...	...

Table: Distance/Similarity between ELTeC Text and OCR Text ; OCR Tool : Kraken

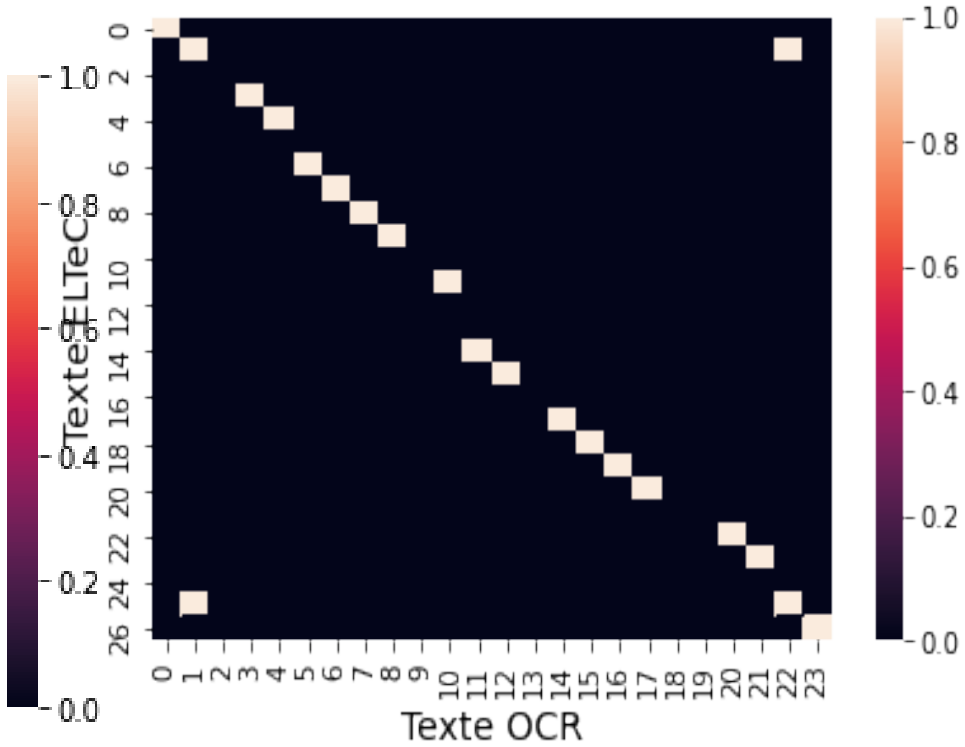
[8] sklearn

Mapping Text and OCR

"Je suis un aventurier" Jacques Dutronc, 1970



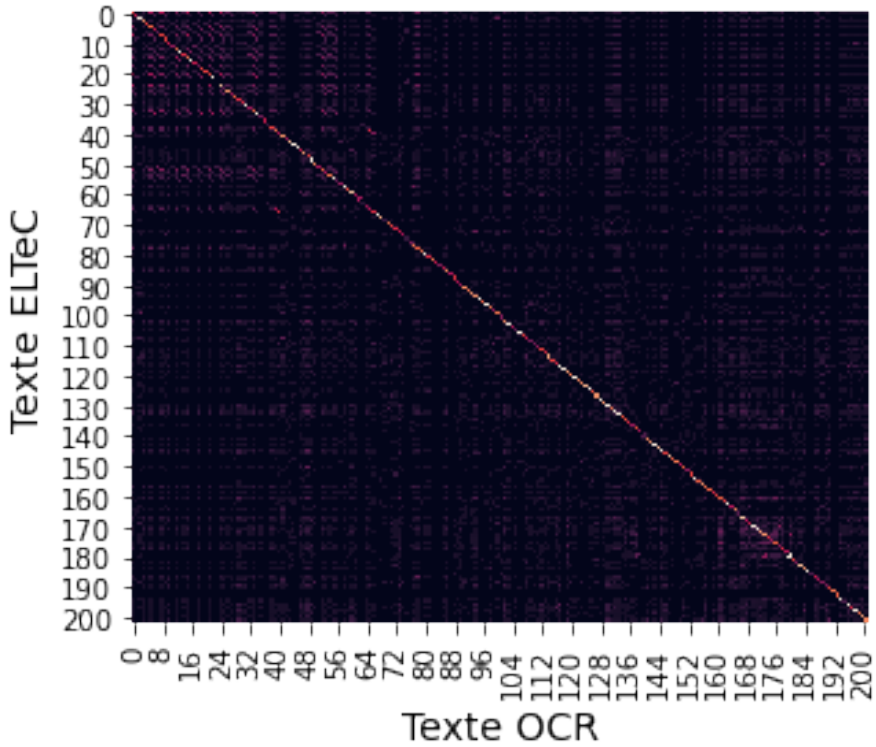
(a) Heatmap representing similarities between the ELTeC corpus text and the OCR output text



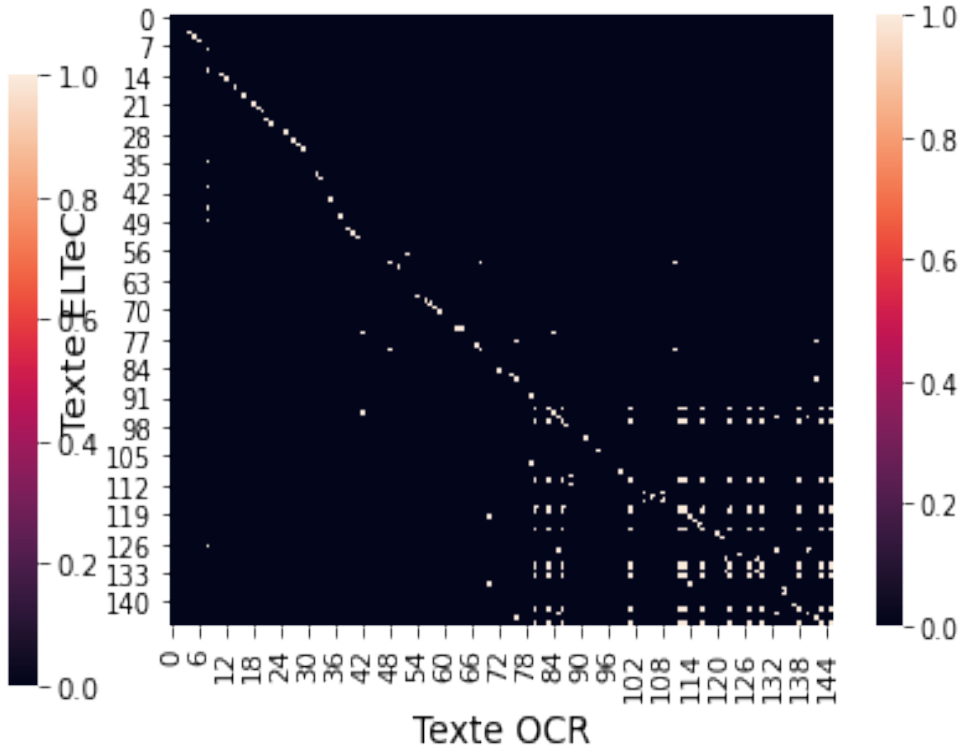
(b) Heatmap representing the similarities between named entities recognized in the ELTeC corpus text and those recognized in the OCR output text.

Mapping Text and OCR

"Dictionnaire géographique", Girault de Saint-Fargeau, (Vol. 1).



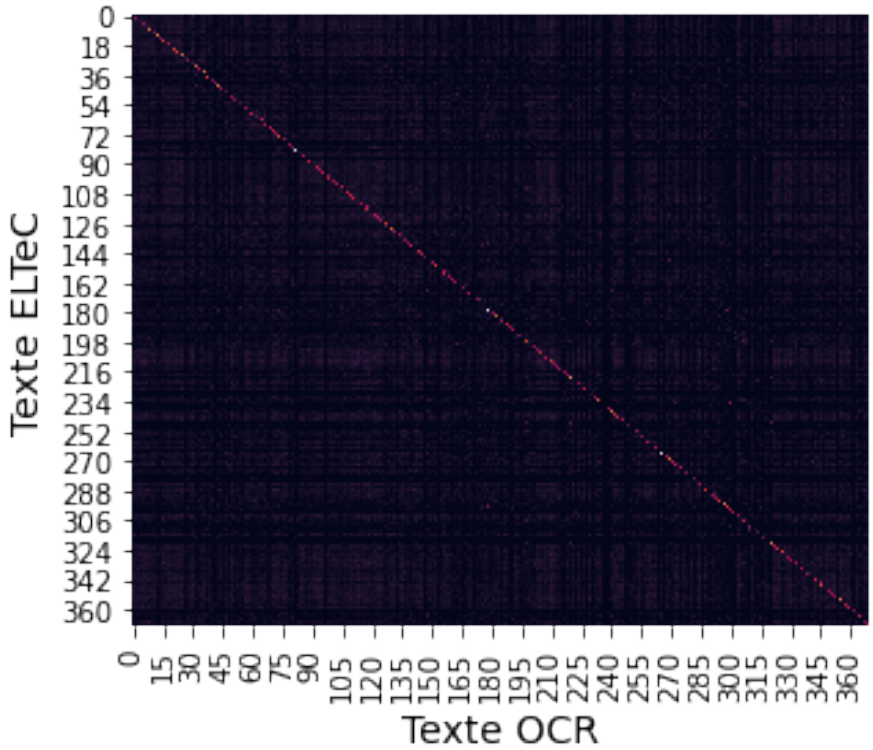
(a) Heatmap representing similarities between the ELTeC corpus text and the OCR output text



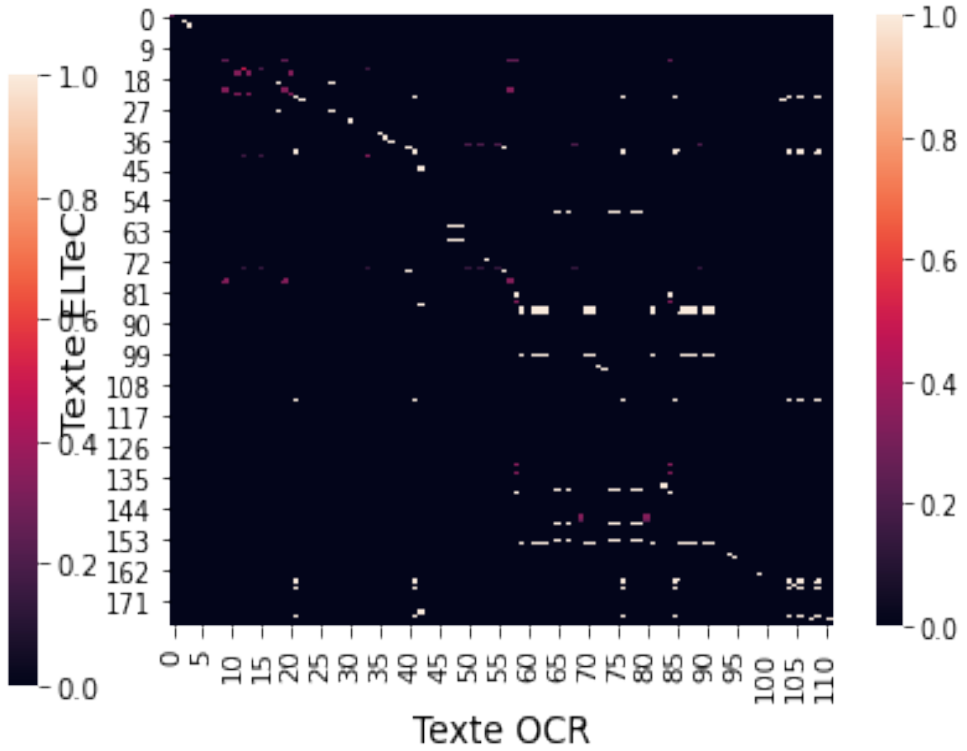
(b) Heatmap representing the similarities between named entities recognized in the ELTeC corpus text and those recognized in the OCR output text.

Mapping Text and OCR

Les trappeurs de l'Arkansas, Gustave Aimard, 1858.



(a) Heatmap representing similarities between the ELTeC corpus text and the OCR output text



(b) Heatmap representing the similarities between named entities recognized in the ELTeC corpus text and those recognized in the OCR output text.

Conclusion and Perspectives

What have learnt so far :

- OCR is the problem ?
- Recurring OCR errors that interfere with NER
- Glass Ceiling in NER [Stanislawek et al., 2019]

Conclusion and Perspectives

What have learnt so far :

- OCR is the problem ?
- Recurring OCR errors that interfere with NER
- Glass Ceiling in NER [Stanislawek et al., 2019]

Some solutions :

- Train models on noisy data ?
- Post-process the NER output rather than input from OCR (parsimonious?)

Conclusion and Perspectives

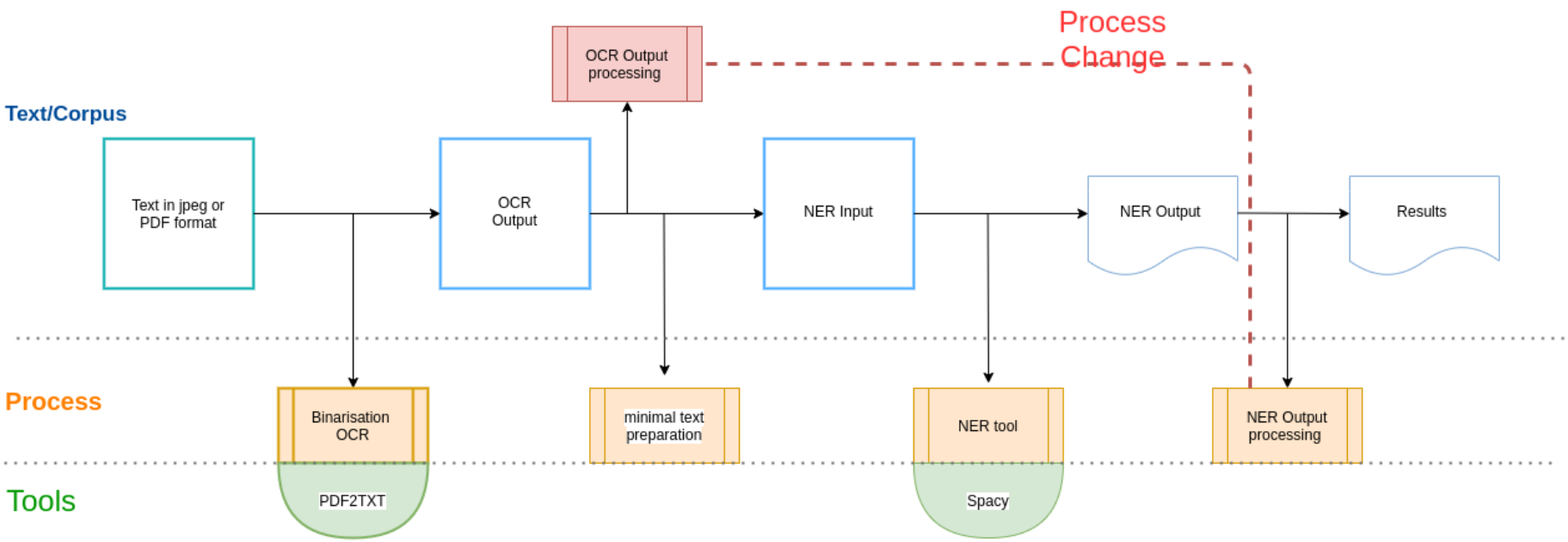


Figure: chaîne de traitement à tester

References I

Stanislawek, T., Wróblewska, A., Wójcicka, A., Ziembicki, D., and Biecek, P. (2019).
Named entity recognition - is there a glass ceiling?
In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 624–633.