Approches quantitatives de textes historiques : quelques (non-)problèmes et comment les aborder

Réunion du groupe de contact FNRS
« Analyse critique et amélioration de la qualité de l'information numérique », ULB
2024-04-16

Simon Hengchen

simon@iguanodon.ai
iguanodon.ai & Université de Genève

Bio

- MA *Langues et littératures germaniques* 2010
- MaSTIC 2012
- PhD STIC 2017

- 2018 - 2020: postdoc COMHIS (Helsinki), groupe d'histoire computationnelle
- 2020 - 2022: postdoc Språkbanken Text (Göteborg), changement sémantique
- 2019 - : chargé d'enseignement Université de Genève
- 2021 - : NLP consulting iguanodon.ai
- 2022 - : steering committee changeiskey.org

Menu du jour

Un OCR de mauvaise qualité est-il une fatalité pour des analyses quantitatives ?

Présentation basée sur du travail partagé avec Mark J. Hill (Kent University) lorsque nous étions tous les deux à Uni Helsinki.

Hill, M.J. and Hengchen, S., 2019. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. Digital Scholarship in the Humanities, 34(4), pp.825-843. https://doi.org/10.1093/llc/fqz024

Postprint en OA:
https://kar.kent.ac.uk/90143/1/Hill_Hengchen_OCR_ECCO_postprint.pdf

OCR et l'analyse quantitative de texte

- Motivation(s):
  - De nombreux articles en DH se plaignent du mauvais OCR (mais ne font rien pour y remédier)
  - De nombreux articles en TAL qui utilisent des données numérisées ne mentionnent jamais l'OCR
  - En DH il semble y avoir dans l'inconscient collectif un rêve utopique d'un futur plus ou moins proche où l'OCR est parfait et "enfin il va être possible de travailler"
  - Il semble y avoir un grand écart dans les attentes entre les fournisseurs de données (“collection-holding institutions”, Wilms 2019) et les chercheurs en ce qui concerne la qualité de l'OCR

OCR et l'analyse quantitative de texte

- Motivation(s):
  - De nombreux articles en DH se plaignent du mauvais OCR (mais ne font rien pour y remédier)
  - De nombreux articles en TAL qui utilisent des données numérisées ne mentionnent jamais l'OCR
  - En DH il semble y avoir dans l'inconscient collectif un rêve utopique d'un futur plus ou moins proche où l'OCR est parfait et "enfin il va être possible de travailler"
  - Il semble y avoir un grand écart dans les attentes entre les fournisseurs de données (“collection-holding institutions”, Wilms 2019) et les chercheurs en ce qui concerne la qualité de l'OCR
- But:
  - Fournir le benchmark définitif et final qui déterminerait une fois pour toutes à quel point un mauvais OCR est trop mauvais

OCR et l'analyse quantitative de texte

- Motivation(s):
  - De nombreux articles en DH se plaignent du mauvais OCR (mais ne font rien pour y remédier)
  - De nombreux articles en TAL qui utilisent des données numérisées ne mentionnent jamais l'OCR
  - En DH il semble y avoir dans l'inconscient collectif un rêve utopique d'un futur plus ou moins proche où l'OCR est parfait et "enfin il va être possible de travailler"
  - Il semble y avoir un grand écart dans les attentes entre les fournisseurs de données (“collection-holding institutions”, Wilms 2019) et les chercheurs en ce qui concerne la qualité de l'OCR
- But plus réaliste:
  - Fournir **un** benchmark qui **aide** les chercheurs à déterminer à quel moment un mauvais OCR est trop mauvais pour le type d’analyse qu’ils souhaitent effectuer (« fitness for use » de Juran)

OCR et l'analyse quantitative de texte

Les travaux antérieurs et ultérieurs sur le sujet comprennent :

- Franzini, G., Kestemont, M., Rotari, G., Jander, M., Ochab, J. K., Franzini, E., Byszuk, J., Rybicki, J. (2018) '**Attributing authorship** in the noisy digitized correspondence of Jacob and Wilhelm Grimm', Frontiers in Digital Humanities, 5(4). DOI: 10.3389/fdigh.2018.00004
- Mutuvi, S., Doucet, A., Odeo, M. and Jatowt, A., 2018, November. Evaluating the impact of OCR errors on **topic modeling**. In International Conference on Asian Digital Libraries (pp. 3-14). Springer, Cham.
- Rodriquez, K.J., Bryant, M., Blanke, T. and Luszczynska, M., 2012. Comparison of **named entity recognition** tools for raw OCR text. In Konvens (pp. 410-414).
- Hamdi, A., Jean-Caurant, A., Sidere, N., Coustaty, M., and Doucet, A. (2019). An Analysis of the Performance of **Named Entity Recognition** over OCRed Documents. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pages 333–334.
- van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B. and Colavizza, G., 2020. Assessing the Impact of OCR Quality on **Downstream NLP Tasks**. In ICAART (1) (pp. 484-496).

Pourquoi ?

1) Les chercheurs passent jusqu'à 80% de leur temps à prétraiter/nettoyer les données et 20% à les analyser (Press, 2016).
2) Les recherches antérieures sur l'impact de l'OCR ne sont pas concluantes :
    - '[T]he current knowledge situation on the users' side as well as on the tool makers' and data providers' side is insufficient and needs to be improved.' (Traub, van Ossenbruggen, and Hardman, 2015)
    - Linguistic DNA a conclu que 'there are too many problems within the OCR dataset to use it' (Hine, 2016).
    - Strange et al (2014): 'Our initial scans had an error rate of 20%... [W]e proceeded to reduce noise… [T]he labour-intensive work of cleaning the data modestly improved the reliability of our test…' MAIS: '**The cleaning was thus desirable but not essential**.'
    - Franzini et al (2018): L'OCR ne doit être propre qu'à environ ~20% pour obtenir des résultats 'higher than chance' en authorship attribution (Handwritten Text Recognition)
    - Eder (2014): propreté d'OCR à 80% n'entraîne pas de résultats significativement pires en authorship attribution

Plan

1. Experimental setup
2. The data
3. Analyses de texte quantitatives
4. Quelques conclusions
5. Limites et pistes à explorer

Experimental setup

Prendre deux corpus constitués de **matériel source identique**, mais dont l'un est océrisé (“sale”, OCR) et l'autre est constitué de données dactylographiées (“propre”, TCP)

- Analyses statistiques:
  a. Exécuter un certain nombre de tests statistiques de base sur les données pour comprendre dans quelle mesure les versions OCR sont bonnes ou mauvaises et diffèrent des données propres.
  b. Diviser les corpus en “quality bands” (groupes de qualité)
- “Text mining” tests:
  a. Faire passer les deux corpus à travers différents outils utilisés en DH pour évaluer leur robustesse par rapport aux données océrisées
  b. (Notons donc que nous n'évaluons pas le résultat d'une analyse particulière, mais plutôt la **différence** entre les résultats.)

The Data

CONTENTS.

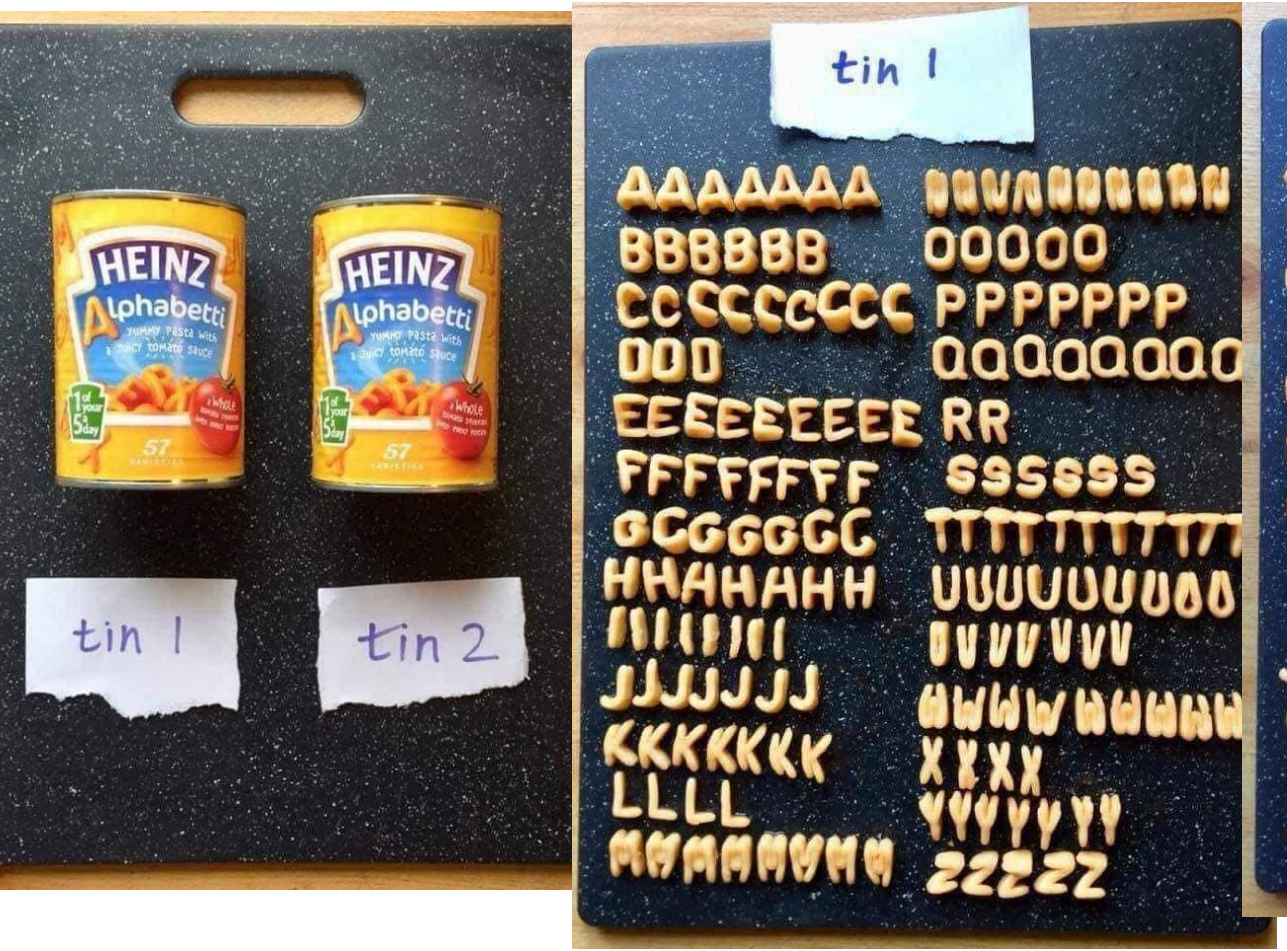

```
61  # CONTENTS.
62
63  C'o N T'E'N. T S.
64
65  C.c~~~~~~~~~~~~
66
67  '. I '
68
69  ''
70
71  CHAP. XIV.
72
73  VISIT the Banks of the GARONNE.-
74  Description of my Country House.-
75  Fall in love with CLAUbINE my
76  Farmer's Daughter.-Account of my
77  (Amour.-Death 'of :.CLAUDINE, and
78  'her interment.-Fati effeats of a bad..
79  Education.m-Reafonsh for being ,so
80  particular in the account .qof. my
81  Amours.-i prefer myself to all ity
82  Countrymen , ..
83
84  CHA P. XV.
85
86  _-. :
87
88  ...*. _
89
90  -_.--
91
92  - '
93
94  - . . ,, J:.''.'/ ' -, ';.
95  My: return to PARIS. x-Summary,. of'
96
97  Events preparatory, t t the Revol-
98  tion.-Patriotifm of the DUyKE of
99
100 ORLEANS..
```

france,\nwritten by himself:\nand translated from the original french,\nby robert ojohnson, esq.\nillustrated with nine engravings.\n - usque adeo permiscuit imis\nlongus summa dies.\nlucan.\nfalso libertatis vocabulum obtendi ab iis, qui privatim degeneres, in publicum exitiosi, nihil spei nisi per discordias habeant.\ntac. an. l. x\n.\nvol. ii.\nlondon: printed for j. debrett, piccadilly. 1794.\n**contents.\nchap. xiv.\ni visit the banks of the garonne. - description of my country house. - fall in love with claudine my farmer's daughter. - account of my amour. - death of claudine, and her interment. - fatal effects of a bad education. - reasons for being so particular in the account of my amours. - i prefer myself to all my countrymen, i\nchap. xv.\nmy return to paris. - summary of events preparatory to the revolution. - patriotism of the duke of orleans.** - advantages of numerous popular assemblies. - flourishing condition of the french republic, 34\nchap. xvi.\nthe duke receives me kindly at paris. - taking of the bastile. - use made of it by the patriots. - real objections to that prison. - delaunay. - berthier, foulon, marat, and i, head bearers. - description of mrs. couteau. - she marches to versailles at the head of five thousand fishwomen. - la fayette. - royal family brought prisoners to paris. - marat, robespierre and i elected members of the convention. - tenth of august, 1792. - patriotism of my mother. - my filial

ECCO vs ECCO-TCP

1. ECCO : Eighteenth Century Collections Online
   a. "Eighteenth Century Collections Online contains over 180,000 titles (200,000 volumes) and more than 32 million pages, making ECCO the premier and irreplaceable resource for eighteenth-century research." (GALE, https://www.gale.com/primary-sources/eighteenth-century-collections-online)
   b. Derrière un paywall
   c. OCR effectué sur des scans de microfilms
2. ECCO-TCP
   a. Subset de ECCO
   b. 2473 livres dactylographiés dans le cadre du projet Text Creation Partnership (TCP)
   c. Open Access

"The Text Creation Partnership was conceived in 1999 between the University of Michigan Library, Bodleian Libraries at the University of Oxford, ProQuest, and the Council on Library and Information Resources (https://textcreationpartnership.org/)"

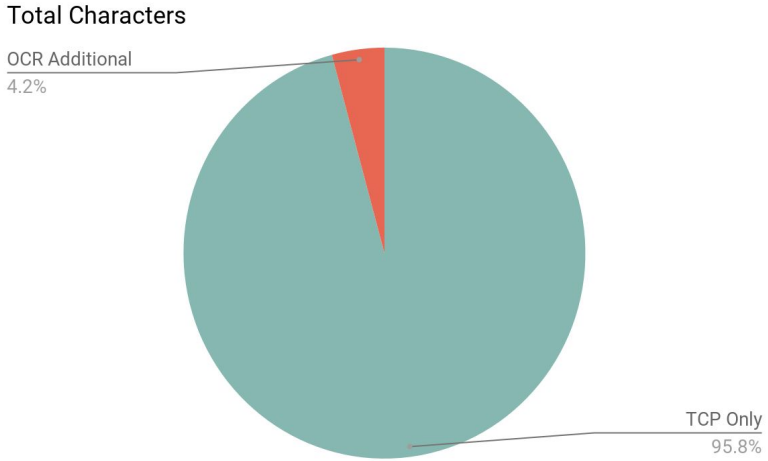


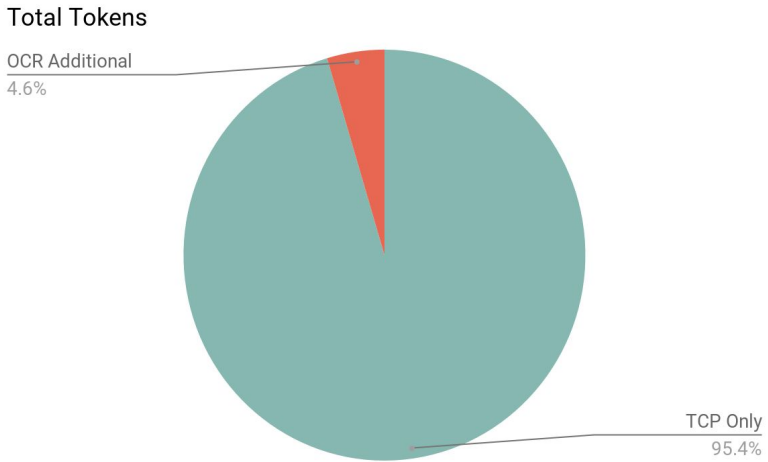
Tailles brutes des deux corpus

En caractères: 394,440,756 (OCR) vs 343,993,778 (TCP)

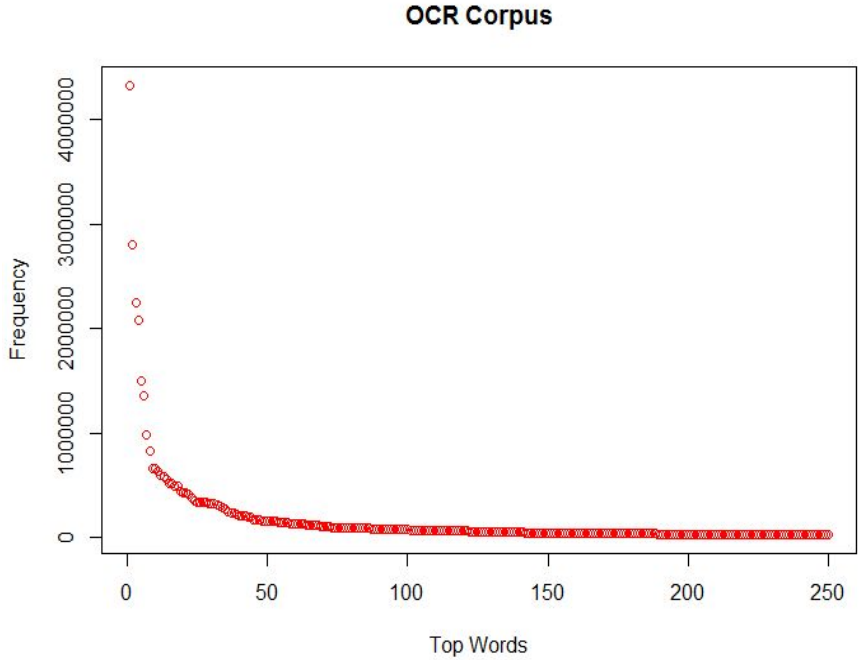
En mots: 95,390,984 (OCR) vs 87,298,605 (TCP)

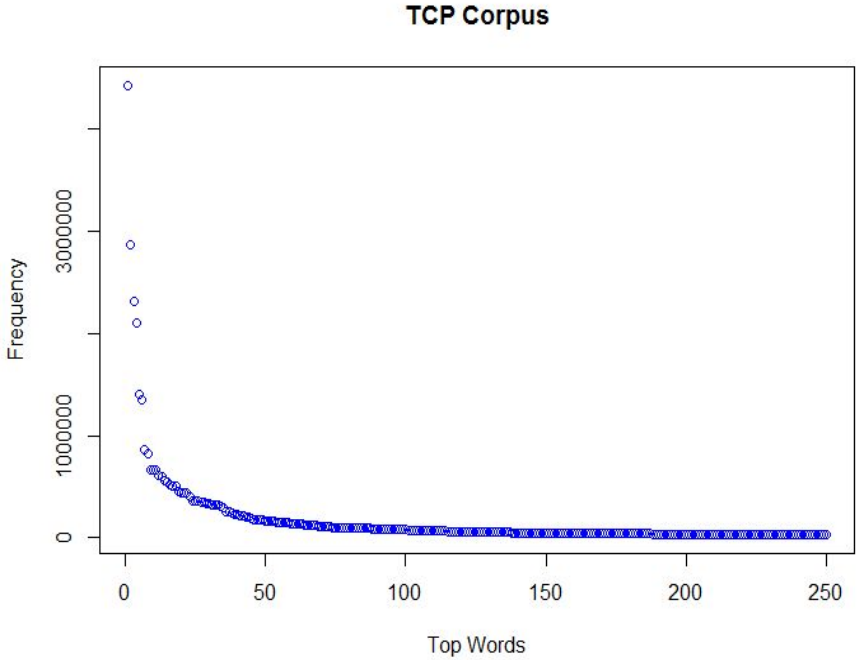
Pour en savoir plus sur la mesure des « erreurs » dans les documents océrisés, voir : Subramaniam et al (2014)





Distribution de Zipf





Les erreurs OCR ne faussent pas la distribution statistique/réelle attendue.

20 mots les plus fréquents

Avec mots-vides (stopwords)

the	the
of	of
and	and
to	to
a	a
in	in
that	i
i	that
his	is
is	it

it	his
with	with
he	he
as	as
for	for
was	be
be	was
by	by
which	which
not	this

Notons que non seulement un mot incorrect dans la liste des mots les plus fréquents indique du bruit supplémentaire (en tant que faux positif), mais il représente également une 'corruption' correspondante (en tant que faux négatif) ailleurs dans le corpus.

Sans mots-vides

time	c
mr	t
sir	mr
little	time
part	e
king	sir
lord	o
life	s
know	fame
s	r

give	p
think	l
c	little
love	de
day	part
people	n
long	d
p	king
found	lord
place	know

Type vs jeton (mot unique vs mot)

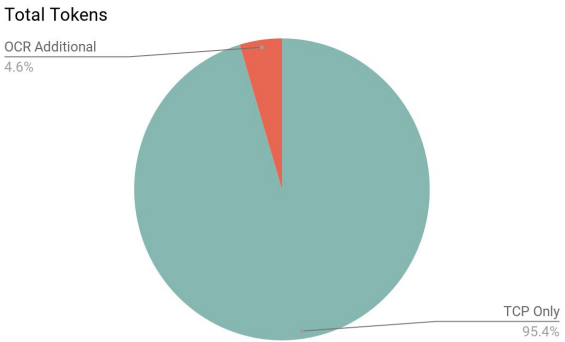
1. Type ("mot unique"):
   a. Classe
   b. != hapax legomenon
   c. Détermine la taille d'un vocabulaire
   d. **Exemple** : chaque entrée d'un dictionnaire représente **un** *type* de la langue française

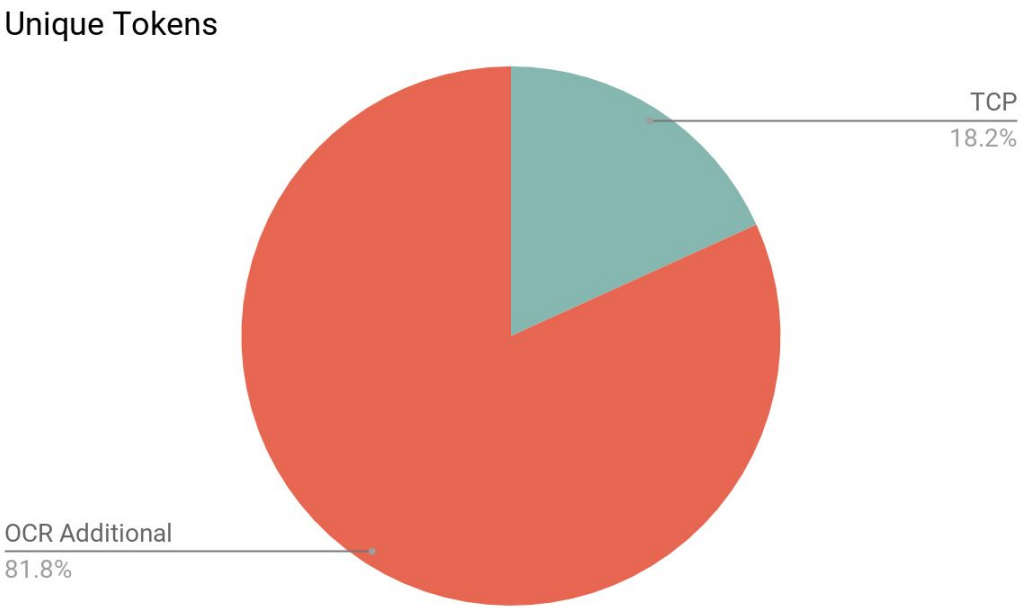
2. Jeton (mot, "token")
   a. Instance d'une classe
   b. **Exemple** : dans le texte *"Il regarde la TV et elle regarde le téléphone"*
      i. 9 jetons / tokens / mots
      ii. 8 types:
         1. Il
         2. regarde (2x)
         3. la
         4. et
         5. …

Représentativité de l'OCR

Mots: 95,390,984 (OCR) vs 87,298,605 (TCP)

Vocabulaire ("types"): 2,703,969 (OCR) vs 765,275 (TCP)

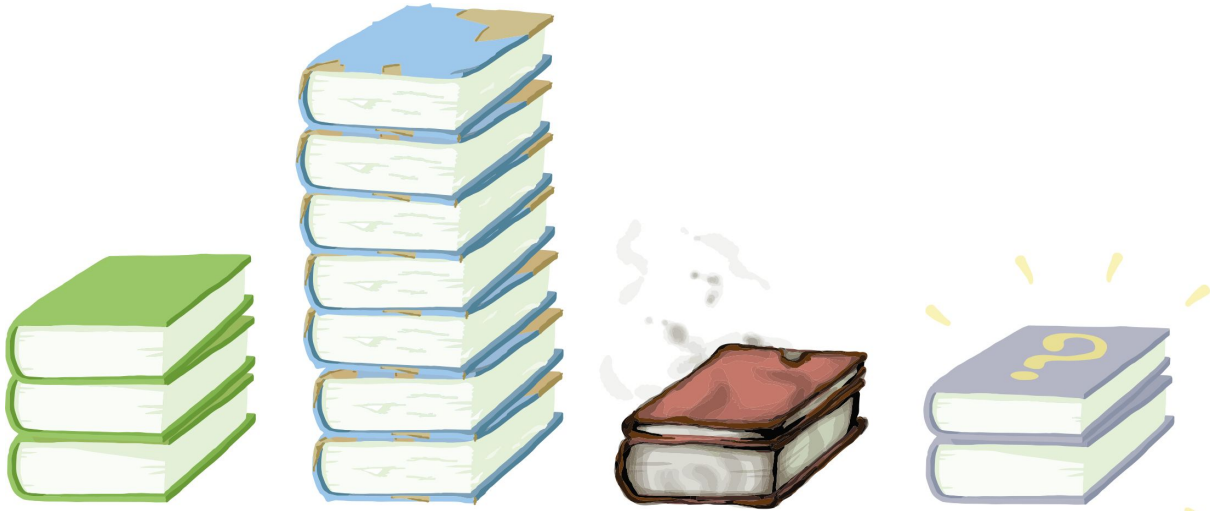




1 mot sur 5 dans le top 500 n’est pas présent dans les deux corpus (sans mots-vides)

"ac" "according" "act" "ad" "afterwards" "answer" "authority" "b" "beauty" "become" "business" "cafe" "case" "character" "com" "con" "conduct" "continued" "dif" "duty" "ed" "effect" "en" "english" "ex" "f" "fall" "fame" "fate" "fay" "fays" "fee" "fide" "fight" "fit" "fix" "foul" "g" "generally" "greatest" "h" "ha" "hall" "happiness" "heaven" "history" "ihe" "ihould" "ill" "immediately" "ing" "interest" "iv" "j" "james" "justice" "k" "kingdom" "la" "laid" "lie" "loft" "lost" "m" "making" "married" "master" "n" "necessary" "object" "obliged" "page" "parliament" "passion" "per" "pro" "purpose" "queen" "r" "re" "reft" "regard" "respect" "rest" "same" "say" "says" "self" "short" "side" "sight" "something" "soul" "strong" "subject" "suppose" "t" "ten" "tion" "u" "un" "used" "vol" "w" "william" "wish" "women" "x" "y" "z"

De quoi ECCO est-il composé ?



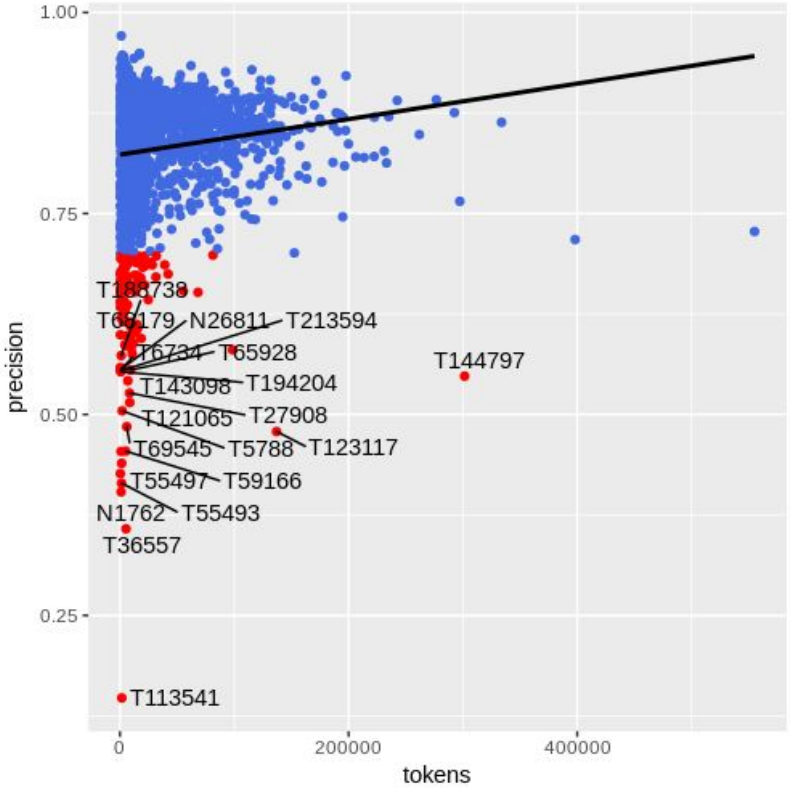
Données correctes

Données que nous pensons +- correctes

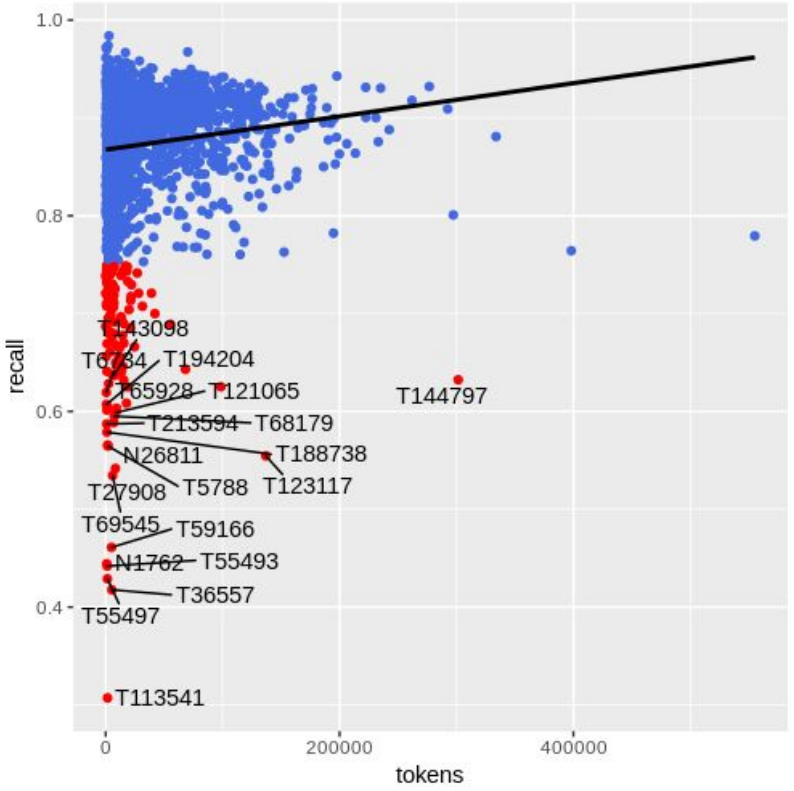
Données manquantes

Données extra

Précision

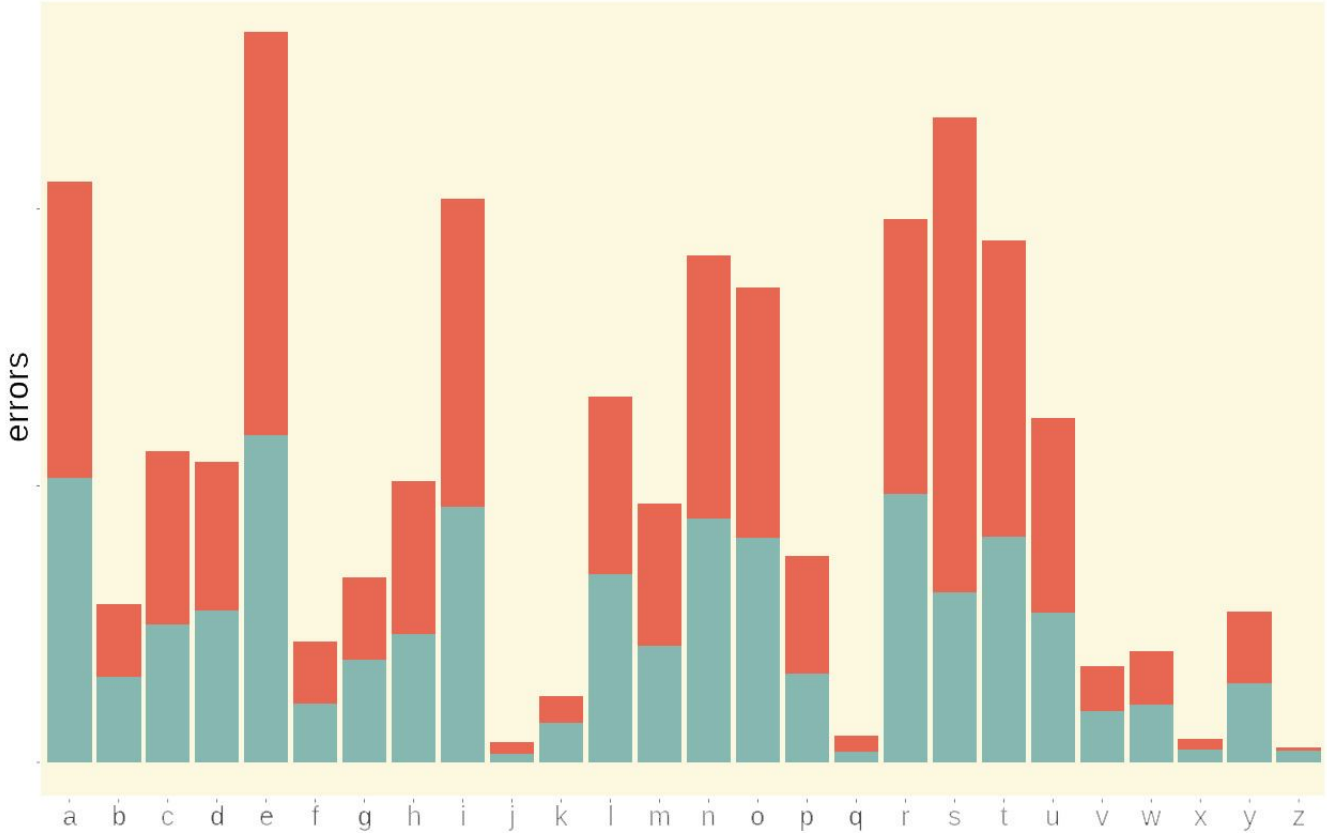


Rappel



Qu’est-ce qui cause les erreurs d’OCR ?

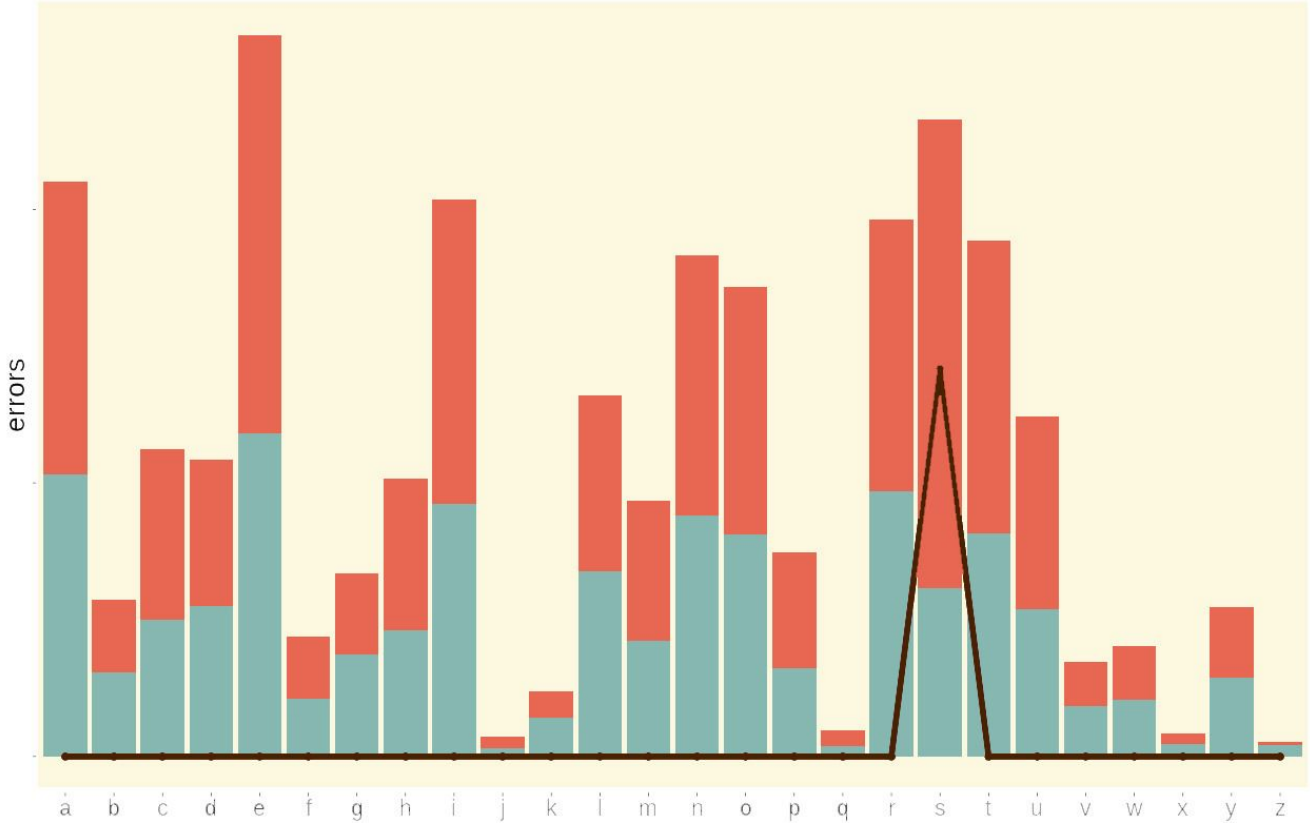
Il y a des variables qui peuvent être inspectées.
Notamment: quelles lettres composent les mots généralement mal océrisés et quel est le nombre moyen de caractères dans ces mots.



Le long-s

Régression binomiale négative tout en contrôlant pour la longueur du mot et des lettres de l'alphabet.

Le seul caractère qui statistiquement pose problème est le « S » ($p < 0,001$).

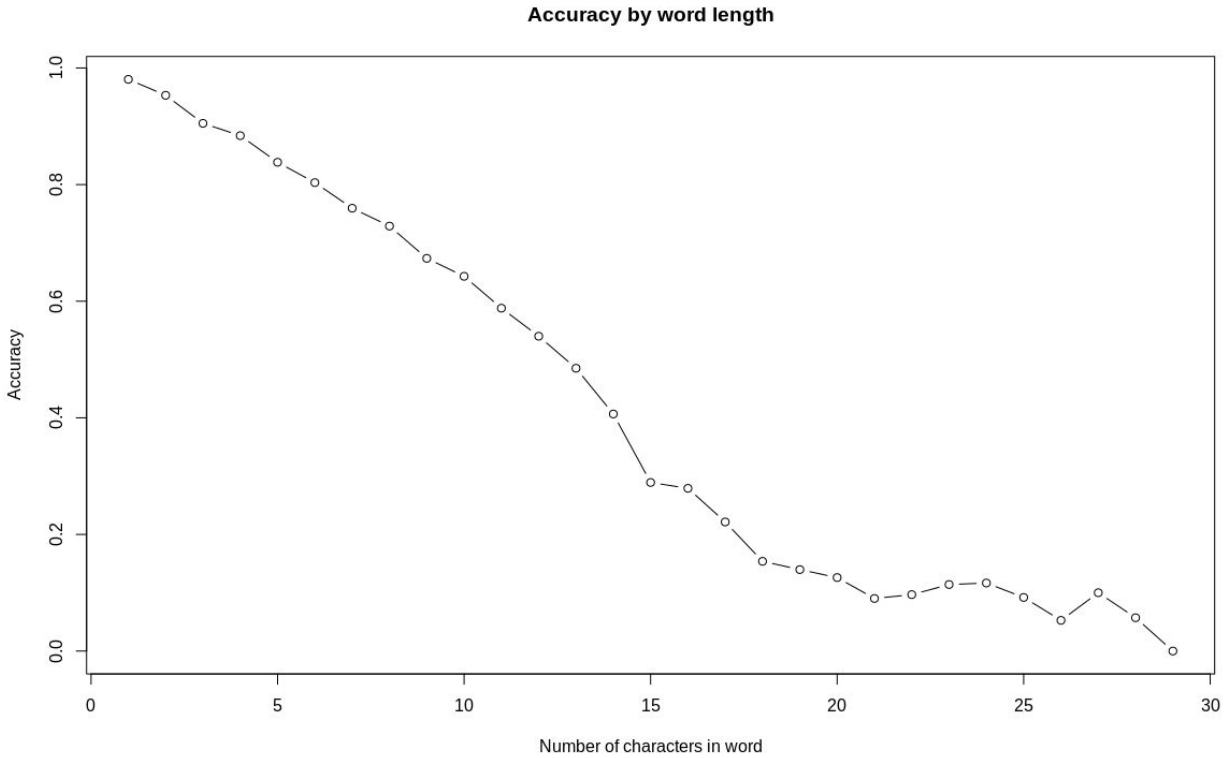


La longueur du mot comme cause des erreurs OCR ?

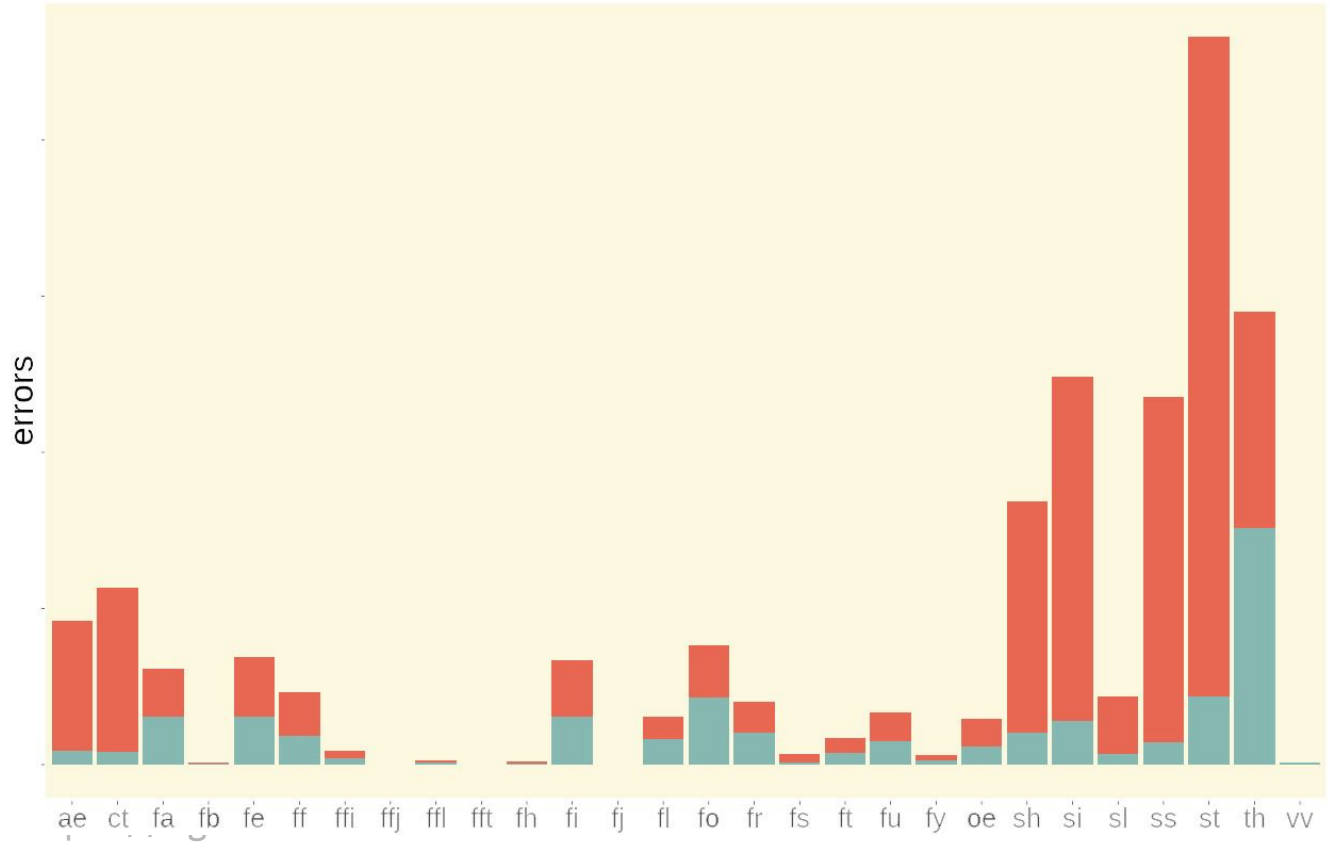
L’exactitude diminue en fonction de la longueur des mots, ce qui est normal.

La longueur du mot en elle-même n’en est pas la cause, mais la composition du mot.

Un mot plus long est plus susceptible de contenir une erreur matérielle ou un caractère/ligature statistiquement problématique pour l’algorithme d’OCR.



Ligatures



st ſt ſt

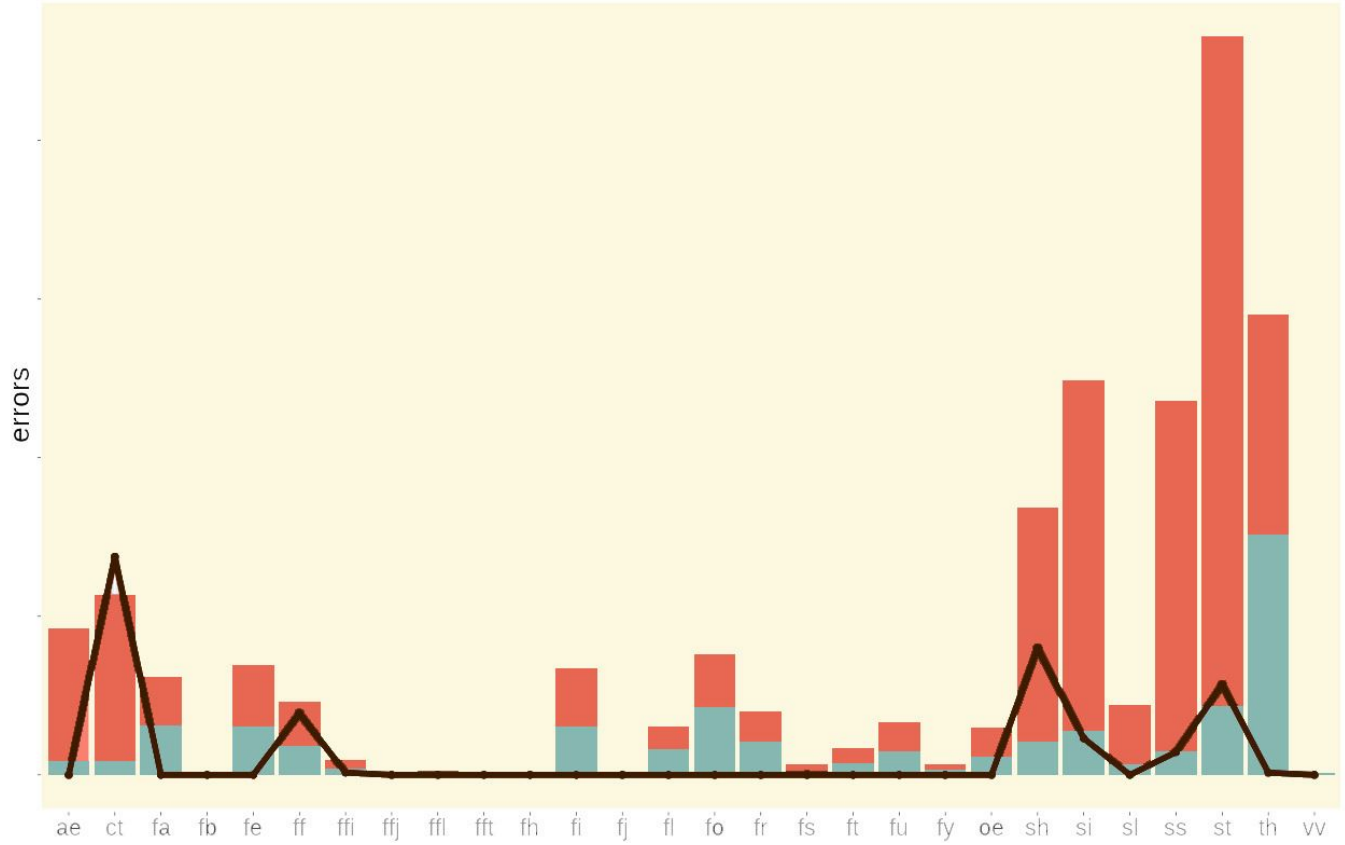
si ſi ſi

ss ſſ ſſ

ff ff ff

ffi ffi ffi

Ligatures



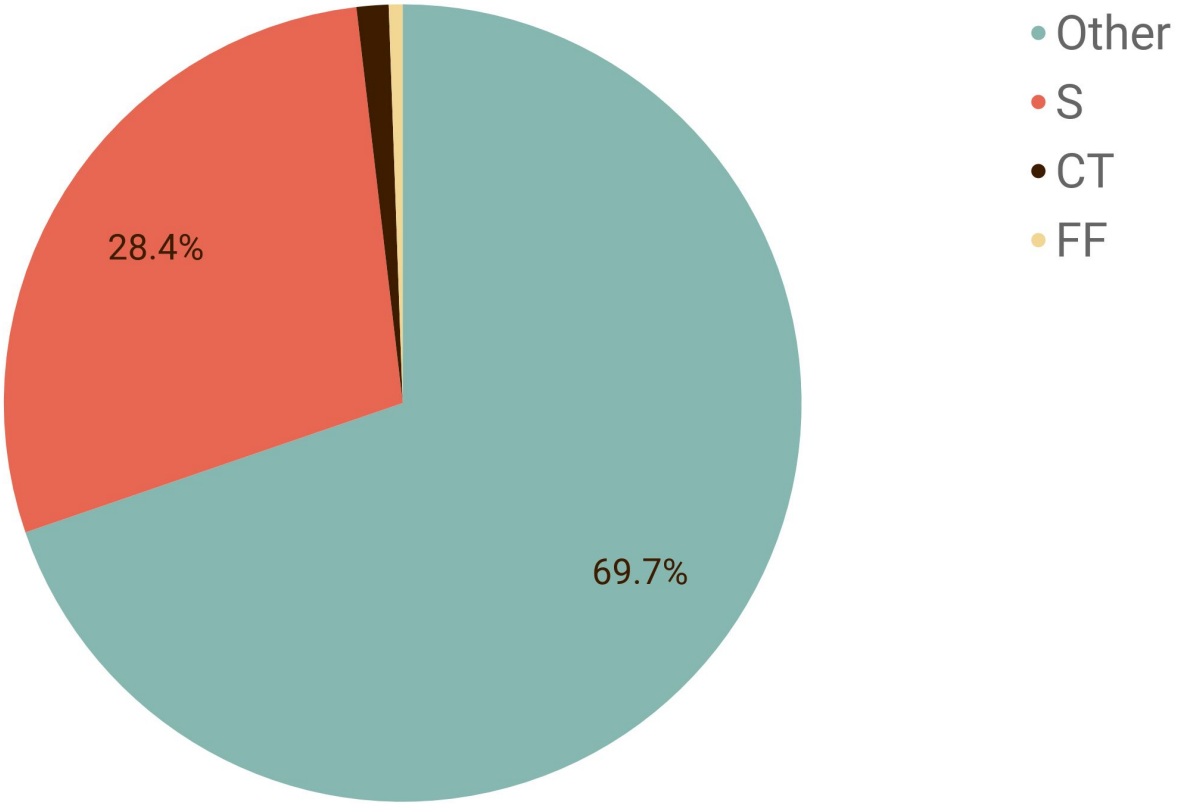
Après avoir contrôlé pour la longueur du mots, la présence de la lettre “s”, et la présence d’autres ligatures.

Les ligatures contenant « s » sont les plus problématiques, mais pas universellement (voir : « fs » et « sl »).

La ligature la plus problématique est « ct » – deux lettres qui n'ont pas été signalées lors du test précédent.

Quelle est l'ampleur du problème ?

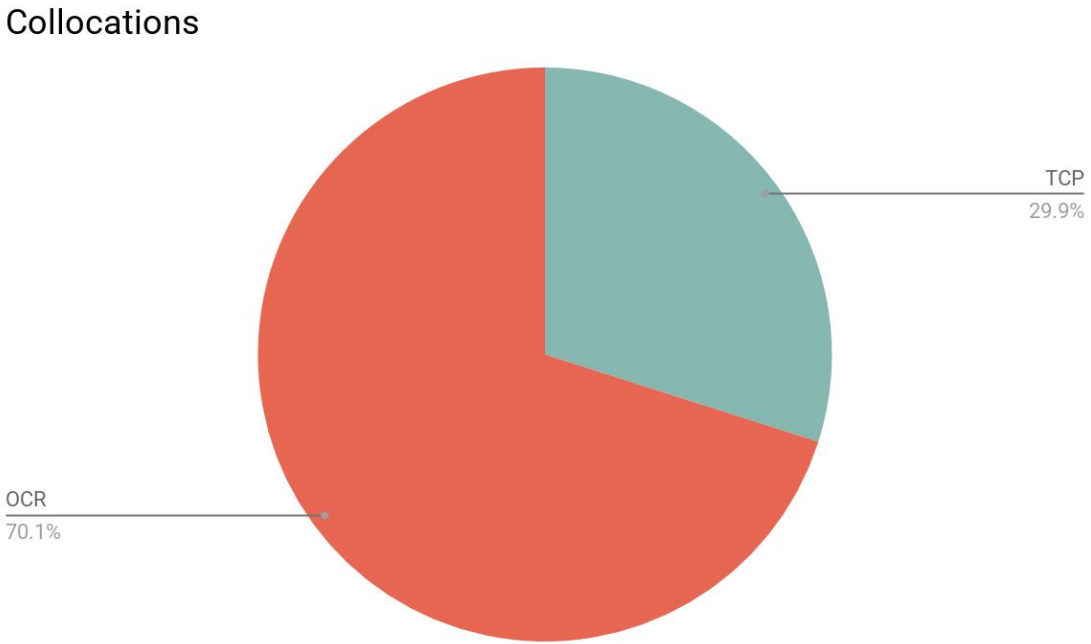
30,26 % de tous les mots du corpus TCP contiennent “S”, “CT”, ou “FF”.



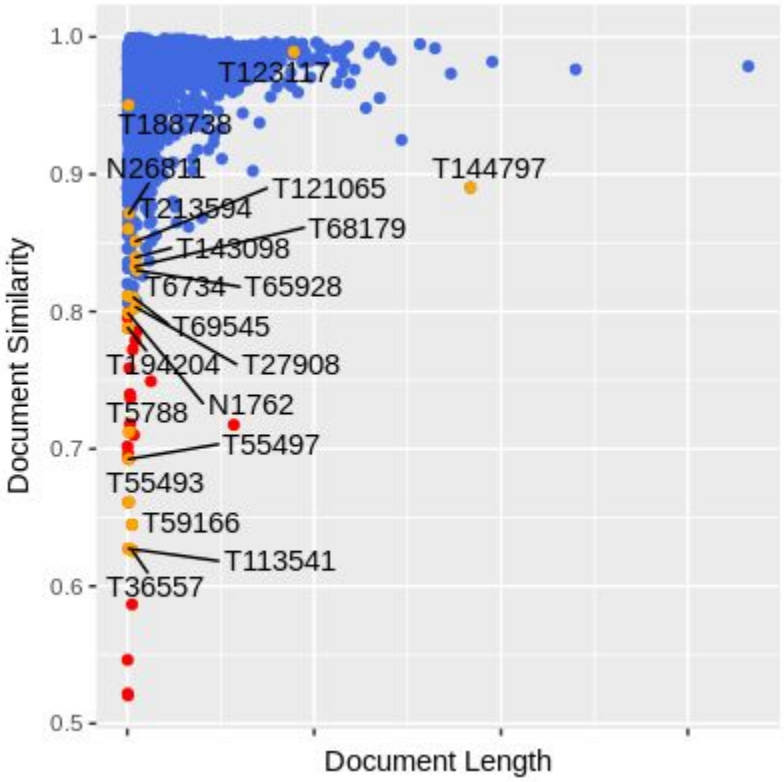
Analyses de texte quantitatives

Collocation

605 569 collocations statistiquement mesurables dans le corpus OCR contre 490 623 dans le corpus TCP (avec un nombre minimum de 10 et aucun mot vide). **319 440 ne correspondent pas.**



Similarité des documents (1-1, OCR vs TCP)



Calcul de la similarité des documents dans l'espace vectoriel, la distance est euclidienne. Comme il s'agit des mêmes documents, la similarité devrait être de 1.

Similarité lexicale - TCP vs OCR Corpus

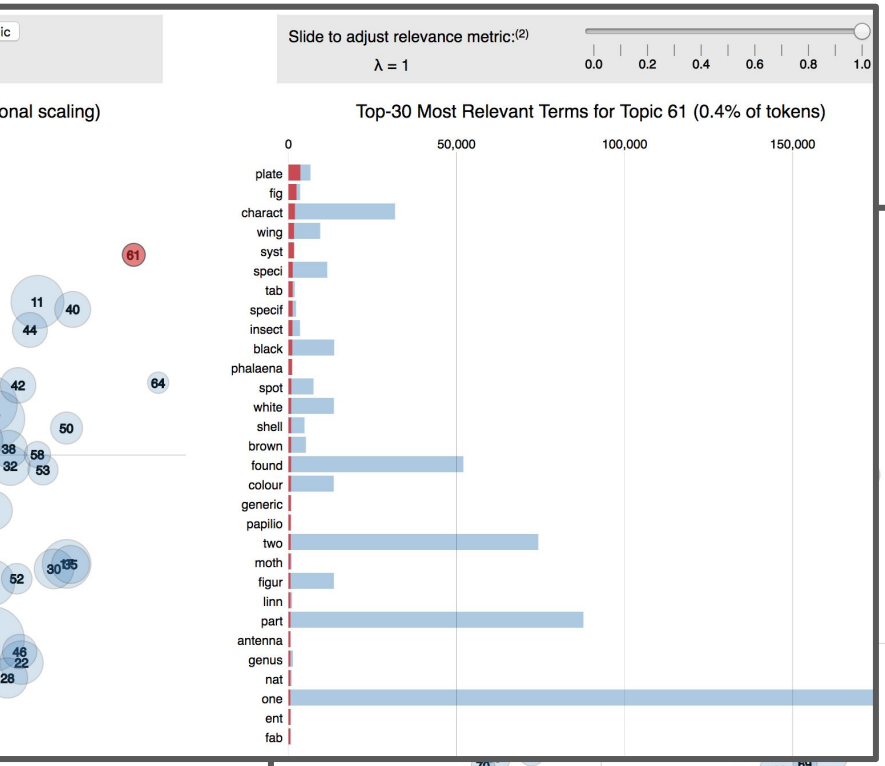
north: south east west side near northern southern places river called
south: north east west near coast side river mile islands sea
east: west south north side near places river sea inhabited called
west: east south north near side river mile places sea coast

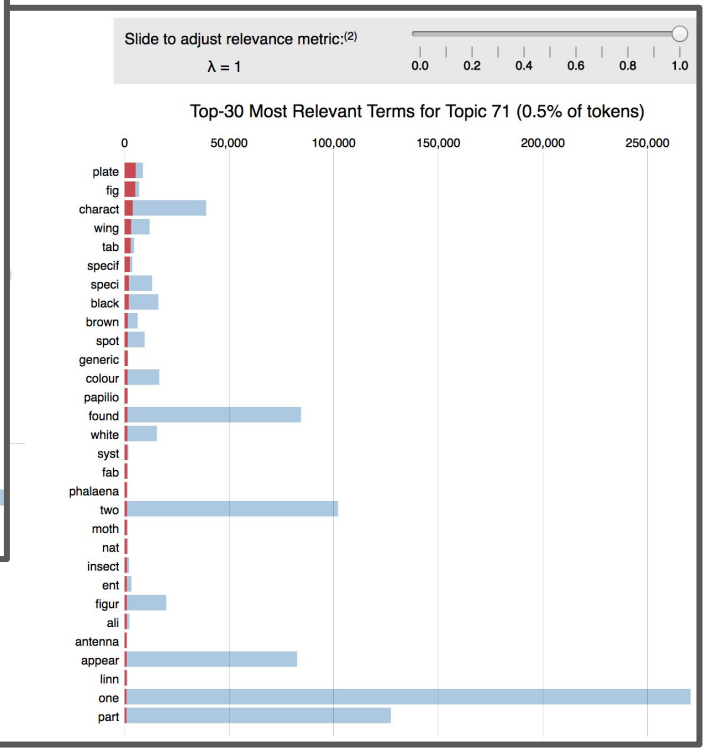
north: south weft east northern southern near places river southeast fide
south: weft north east coast near river mile inhabited bay islands
east: weft south north near river sea inhabited places coast eastern
west: osne thebreadth addingham plxnii eisto 2 iiiz1i stvini statelian felftevidentlygiright faihionriable

weft: east south north river near coast sea welt mile inhabited

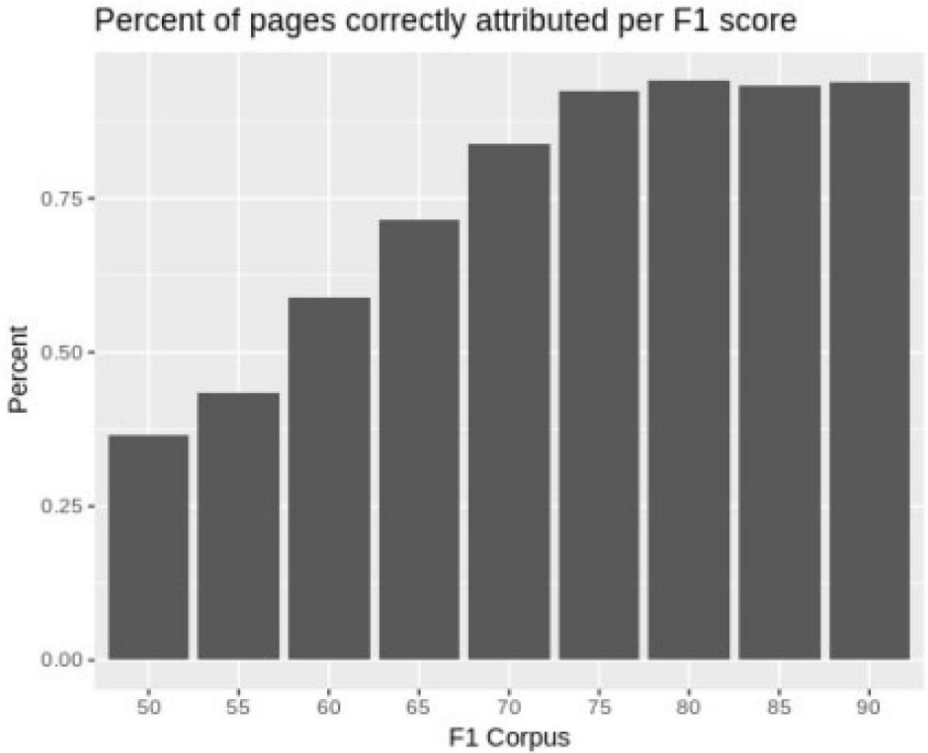
Topic Modelling

- STM + seed pour trouver le K “optimal”.
- Ks légèrement différents, mais résultats très similaires.
- Ex : Topics 61/71, « insectes », distributions de mots extrêmement similaires





Authorship attribution



- Top 25 des auteurs les plus prolifiques
- 3 tests (delta, k-nearest neighbour, nearest centroid classifier)
- 3 'features' différentes (unigrams, bigrams, trigrams)
- Testé avec les n features les plus fréquentes, n dans [100, 200, 300, 400, 500]

Conclusions

1. Les algorithmes testés sont règle générale assez robustes contre l'OCR
   a. … dans un contexte "sac de mots" (bag-of-words), mais pas que
   b. … dans contexte quantitatif, et moins dans un contexte nuancé (changement syntaxique, généalogie, etc.)
2. Les erreurs OCR ne sont pas, dans l'ensemble, aléatoires et ne doivent donc pas être traitées comme telles dans une analyse.
3. Une qualité de 80 % semble être un seuil décent pour la plupart des tâches quantitatives, **la taille du corpus est souvent un paramètre plus important que la qualité du corpus**

Limites et pistes d'amélioration

1. Tests effectués sur un corpus avec :
    a. Un type d'OCR,
        i. sur **une** police d'écriture
        ii. avec un type d'algorithme d'OCR
        iii. … vieillot
    b. Une seule langue

2. Les analyses effectuées, bien que représentant une large partie des analyses utilisées en DH, ne sont plus à la pointe du TAL
    a. Une tokénisation en "sub-words" est souvent préférée
    b. D'autres manières de vectoriser des textes existent et sont souvent préférées

Outils utilisés pour les analyses

Benoit, K. (2018). *quanteda: Quantitative Analysis of Textual Data*. R package version 1.3.0. http://quanteda.io.

Michalke, M. (2017). *koRpus: An R Package for Text Analysis* (Version 0.10-2). https://reaktanz.de/?c=hacking&s=koRpus

Margaret E. Roberts, Brandon M. Stewart and Dustin Tingley (2018). *stm: R Package for Structural Topic Models.* http://www.structuraltopicmodel.com.

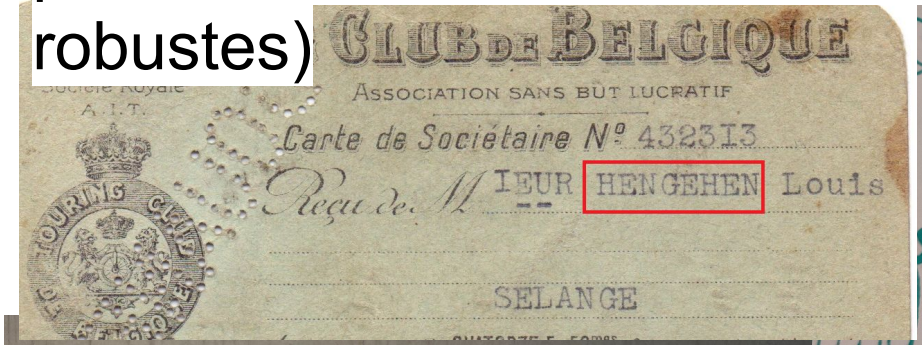
Carson Sievert and Kenny Shirley (2014). “LDAvis: A method for visualizing and interpreting topics.” *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. http://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf.

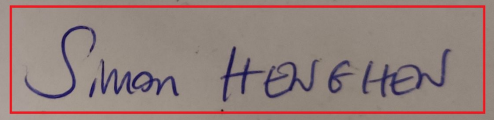
Eder, M., Rybicki, J. and Kestemont, M. (2016). “Stylometry with R: a package for computational text analysis.” *R Journal*, 8(1): 107-121. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

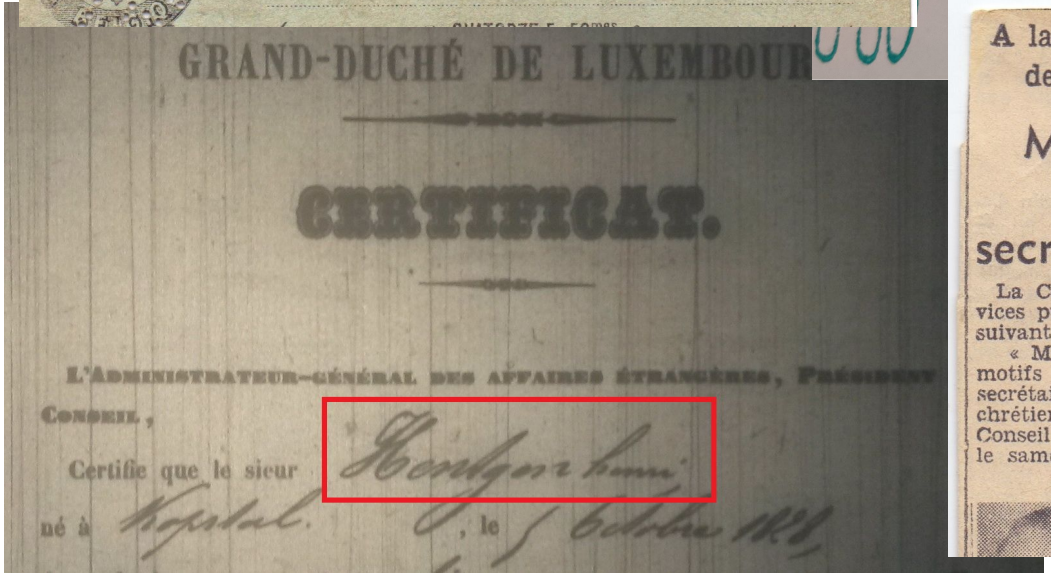
Remerciements

The Comhis Collective: Ali Ijaz, Antti Kanner, Leo Lahti, Eetu Mäkelä, Jani Marjanen, Hege Roivainen, Tanja Säily, Iiro Tiihonen, Mikko Tolonen, Ville Vaara. Also: Adrienne Hawkes, Jack Cunliffe, Johan Ahlback, Giovanni Colavizza

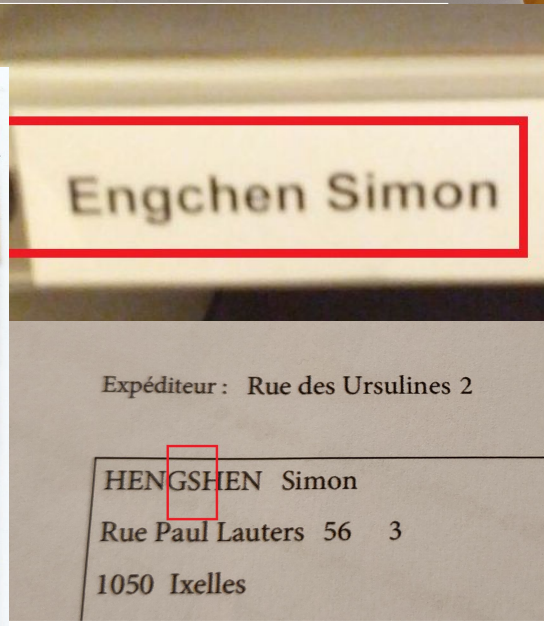
(un OCR parfait ne résoudra pas non plus tous les problèmes, ce dont nous avons besoin ce sont des modèles robustes)


CLUB DE BELGIQUE
ASSOCIATION SANS BUT LUCRATIF
Carte de Sociétaire N° 432313
Reçu de M IEUR HENGEHEN Louis
SELANGE


Simon HENGHEN


GRAND-DUCHÉ DE LUXEMBOU
CERTIFICAT.
L'ADMINISTRATEUR-GÉNÉRAL DES AFFAIRES ÉTRANGÈRES, PRÉSIDENT CONSEIL,
Certifie que le sieur Hengchen
né à Kopstal, le 5 Octobre 1891


A la Centrale chrétienne des Services publics
M. Heugchen est élu secrétaire général
La Centrale chrétienne des Services publics nous adresse le texte suivant :
« M. Buyens ayant remis, pour motifs de santé, sa démission de secrétaire général de la Centrale chrétienne des Services publics, le Conseil de la Centrale s'est réuni le samedi 19 octobre, à Bruxelles.
Il a élu secrétaire général M. Armand Heugchen, ancien secrétaire


Engchen Simon
Expéditeur : Rue des Ursulines 2
HENGSHEN Simon
Rue Paul Lauters 56 3
1050 Ixelles

this slide intentionally left blank

this slide intentionally left blank