Rapprocher des éléments textuels similaires : gestion des sauts qualitatifs en général et de la variation morphologique en particulier

Gaël Lejeune, gael.lejeune@sorbonne-universite.fr

26 octobre 2023

Sorbonne Université

Define : saut qualitatif

Différentes variations peuvent affecter les données textuelles,

Define : saut qualitatif

Différentes variations peuvent affecter les données textuelles,

certaines sont légères et ne sont pas étonanntes par rapport à la moyenne (de l'attendu)

Define : saut qualitatif

Différentes variations peuvent affecter les données textuelles,

certaines sont légères et ne sont pas étonanntes par rapport à la moyenne (de l'attendu)

, d'autres non : ce sont les **sauts**.

Define : saut qualitatif

Différentes variations peuvent affecter les données textuelles,

certaines sont légères et ne sont pas étonanntes par rapport à la moyenne (de l'attendu)

, d'autres non : ce sont les **sauts**.

Certains de ces sauts sont liés à des propriétés de sous-corpus (langue, genre, format . . .)

Define : saut qualitatif

Différentes variations peuvent affecter les données textuelles,

certaines sont légères et ne sont pas étonanntes par rapport à la moyenne (de l'attendu)

, d'autres non : ce sont les **sauts**.

Certains de ces sauts sont liés à des propriétés de sous-corpus (langue, genre, format ...)

d'autres sont d'ordre qualitatif :

Define : saut qualitatif

Différentes variations peuvent affecter les données textuelles,

certaines sont légères et ne sont pas étonanntes par rapport à la moyenne (de l'attendu)

, d'autres non : ce sont les **sauts**.

Certains de ces sauts sont liés à des propriétés de sous-corpus (langue, genre, format ...)

d'autres sont d'ordre qualitatif :

- Segments surnuméraires (OCR et Web Scraping)
- Variation interne à un segment (bruit et silence)
- Perte d'informations :
  - sur la structure du document
  - méta-données lacunaires, disparates ...

Utilisabilité de chaque document du corpus ?

Comparer deux résultats de scraping

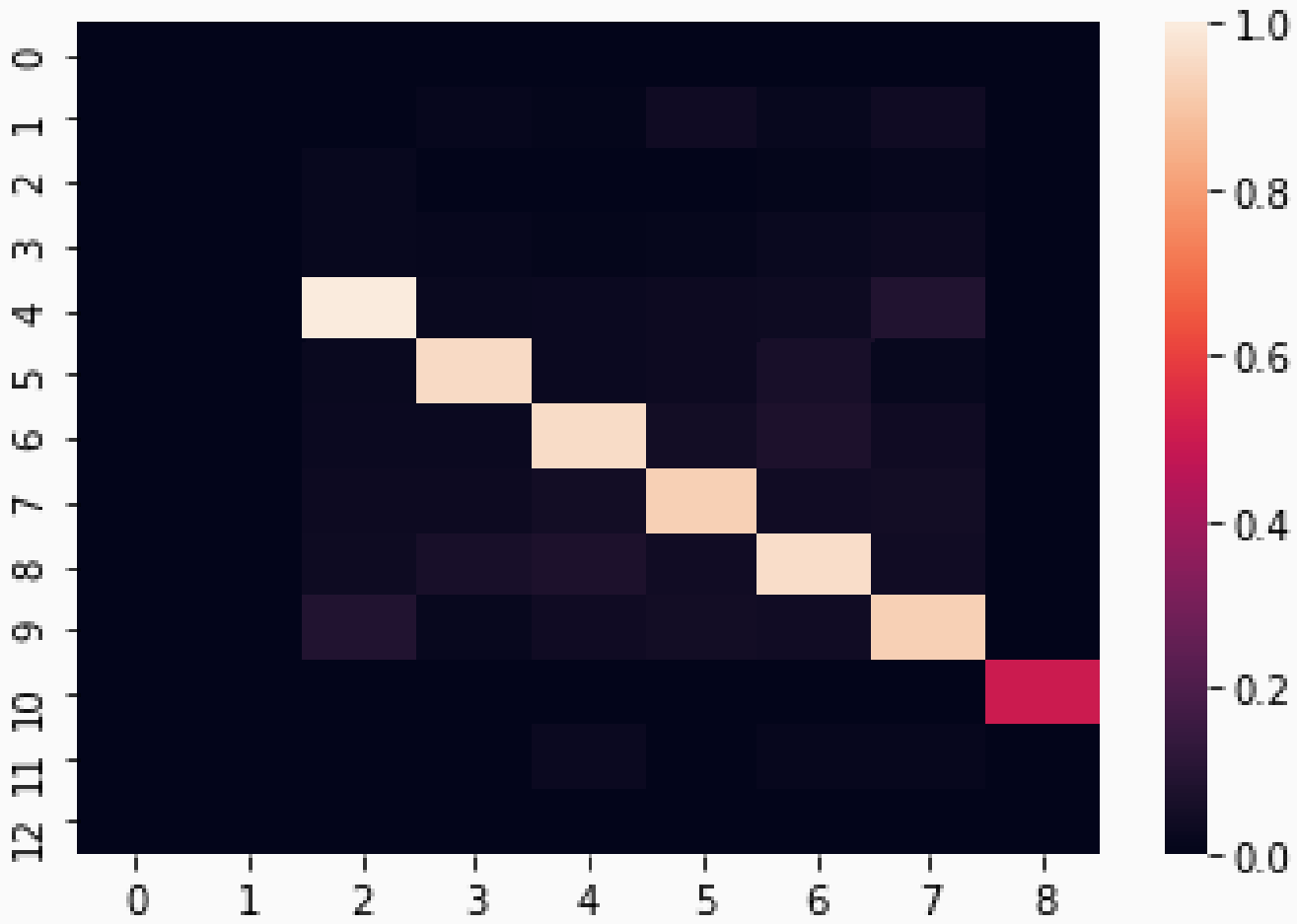


Figure 1 – Carte de chaleur (*Heatmap*)

Comment repérer les sauts qualitatifs

- Faire varier les observables (Tweets, Ghoul et Lejeune)
- **Analyser à différents grains : page**, paragraphe (Ex : Mazarinades, Abiven & Lejeune)

Comment repérer les sauts qualitatifs

- Faire varier les observables (Tweets, Ghoul et Lejeune)
- **Analyser à différents grains : page**, paragraphe (Ex : Mazarinades, Abiven & Lejeune)
- Ne pas trop généraliser un bon (ou mauvais) résultat (Ex : scraping, Barbaresi & Lejeune)

Comment repérer les sauts qualitatifs

- Faire varier les observables (Tweets, Ghoul et Lejeune)
- **Analyser à différents grains : page**, paragraphe (Ex : Mazarinades, Abiven & Lejeune)
- Ne pas trop généraliser un bon (ou mauvais) résultat (Ex : scraping, Barbaresi & Lejeune)
- **Exploiter des mesures non-supervisées d'évaluation** (Ex : OCR, Tanguy)

Comment repérer les sauts qualitatifs

- Faire varier les observables (Tweets, Ghoul et Lejeune)
- **Analyser à différents grains : page**, paragraphe (Ex : Mazarinades, Abiven & Lejeune)
- Ne pas trop généraliser un bon (ou mauvais) résultat (Ex : scraping, Barbaresi & Lejeune)
- **Exploiter des mesures non-supervisées d'évaluation** (Ex : OCR, Tanguy)
- Isoler les parties malades d'un corpus (Ex : OCR + NER, Parfait et Lejeune)

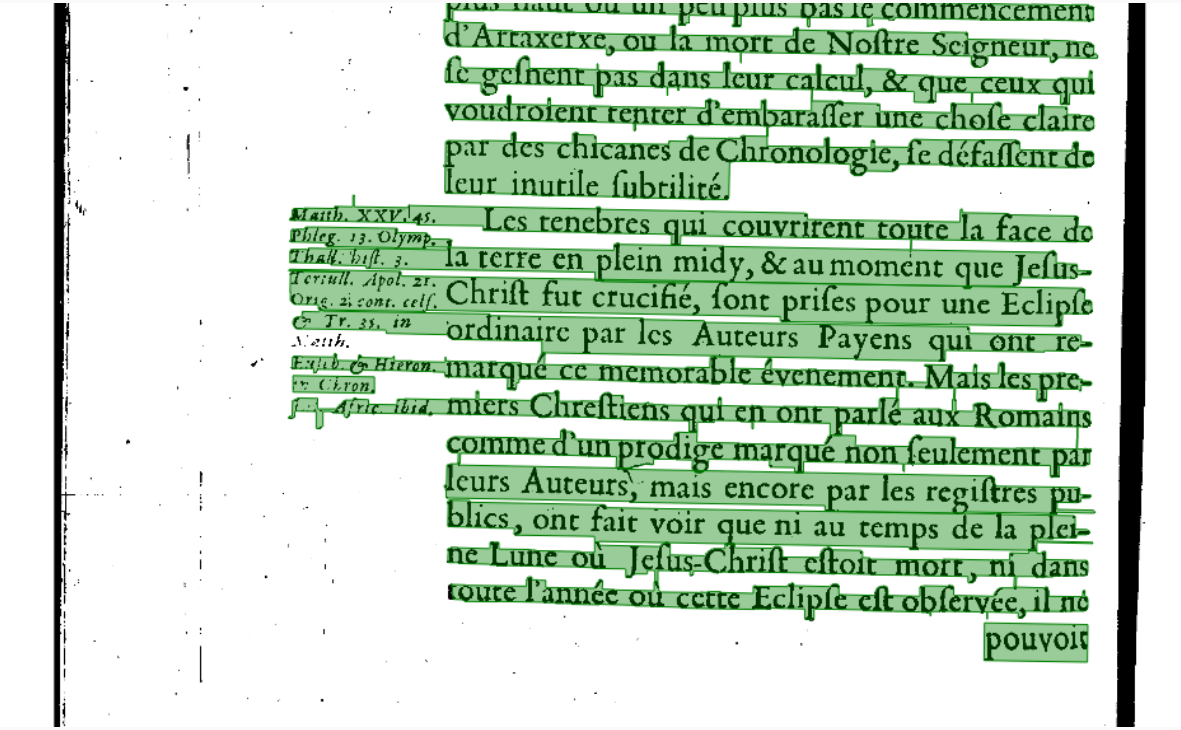
Comment repérer les sauts qualitatifs

- Faire varier les observables (Tweets, Ghoul et Lejeune)
- **Analyser à différents grains : page**, paragraphe (Ex : Mazarinades, Abiven & Lejeune)
- Ne pas trop généraliser un bon (ou mauvais) résultat (Ex : scraping, Barbaresi & Lejeune)
- **Exploiter des mesures non-supervisées d'évaluation** (Ex : OCR, Tanguy)
- Isoler les parties malades d'un corpus (Ex : OCR + NER, Parfait et Lejeune)
- Passer du seuil à la vraisemblance (Ex : alignement de tweets, Bezançon et Lejeune)

Comment repérer les sauts qualitatifs

- Faire varier les observables (Tweets, Ghoul et Lejeune)
- **Analyser à différents grains : page**, paragraphe (Ex : Mazarinades, Abiven & Lejeune)
- Ne pas trop généraliser un bon (ou mauvais) résultat (Ex : scraping, Barbaresi & Lejeune)
- **Exploiter des mesures non-supervisées d'évaluation** (Ex : OCR, Tanguy)
- Isoler les parties malades d'un corpus (Ex : OCR + NER, Parfait et Lejeune)
- Passer du seuil à la vraisemblance (Ex : alignement de tweets, Bezançon et Lejeune)
- **Combiner différentes "versions d'un texte"**, reconstruction ou anastylose (Ex : Chansons)

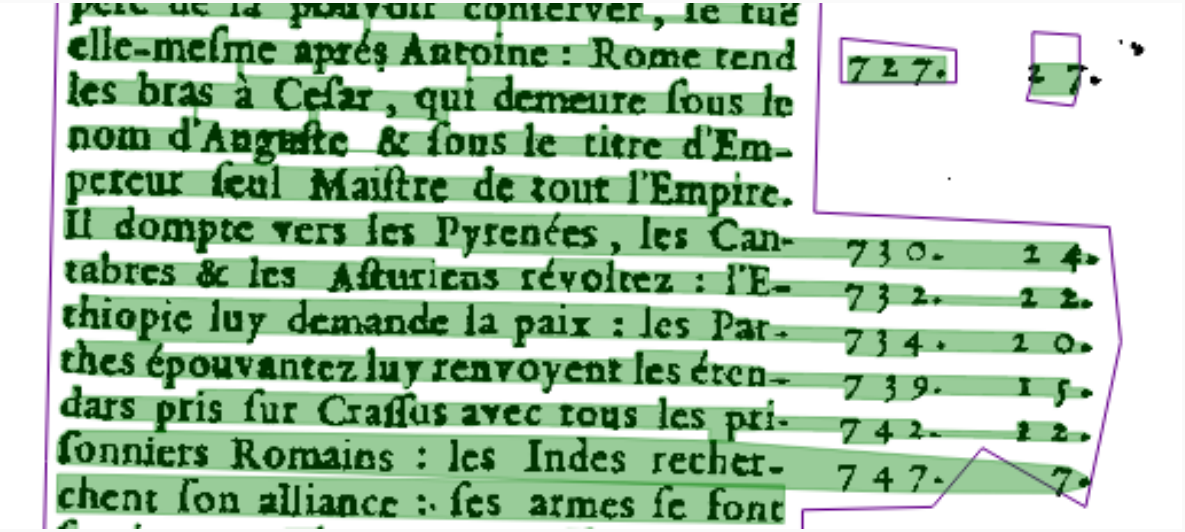
Focus sur les Documents anciens (I)



plus haut ou un peu plus bas le commencement
d'Artaxerxe, ou la mort de Noſtre Seigneur, ne
ſe geſnent pas dans leur calcul, & que ceux qui
voudroient tenter d'embaraſſer une choſe claire
par des chicanes de Chronologie, ſe défaſſent de
leur inutile ſubtilité.

Matth. XXV. 45. Les tenebres qui couvrirent toute la face de
Phleg. 13. Olymp.
Thall. hiſt. 3. la terre en plein midy, & au moment que Jeſus-
Tertull. Apol. 21.
Orig. 2. cont. celſ. Chriſt fut crucifié, ſont priſes pour une Eclipſe
& Tr. 35. in Matth. ordinaire par les Auteurs Payens qui ont re-
Euſeb. & Hieron. in Chron. marqué ce memorable évenement. Mais les pre-
Sext. Afric. ibid. miers Chreſtiens qui en ont parlé aux Romains
comme d'un prodige marqué non ſeulement par
leurs Auteurs, mais encore par les regiſtres pu-
blics, ont fait voir que ni au temps de la plei-
ne Lune où Jeſus-Chriſt eſtoit mort, ni dans
toute l'année où cette Eclipſe eſt obſervée, il ne
pouvoit

Focus sur les Documents anciens (II)



perc de la pouvoir conſerver, le tue
elle-meſme aprés Antoine : Rome tend
les bras à Ceſar, qui demeure ſous le
nom d'Auguſte & ſous le titre d'Em-
pereur ſeul Maiſtre de tout l'Empire.
Il dompte vers les Pyrenées, les Can-
tabres & les Aſturiens révoltez : l'E-
thiopie luy demande la paix : les Par-
thes épouvantez luy renvoyent les éten-
dars pris ſur Craſſus avec tous les pri-
ſonniers Romains : les Indes recher-
chent ſon alliance : ſes armes ſe font

727. 27.
730. 24.
732. 22.
734. 20.
739. 15.
742. 12.
747. 7.

Comment mesurer l'étendue des dégats ? (I)

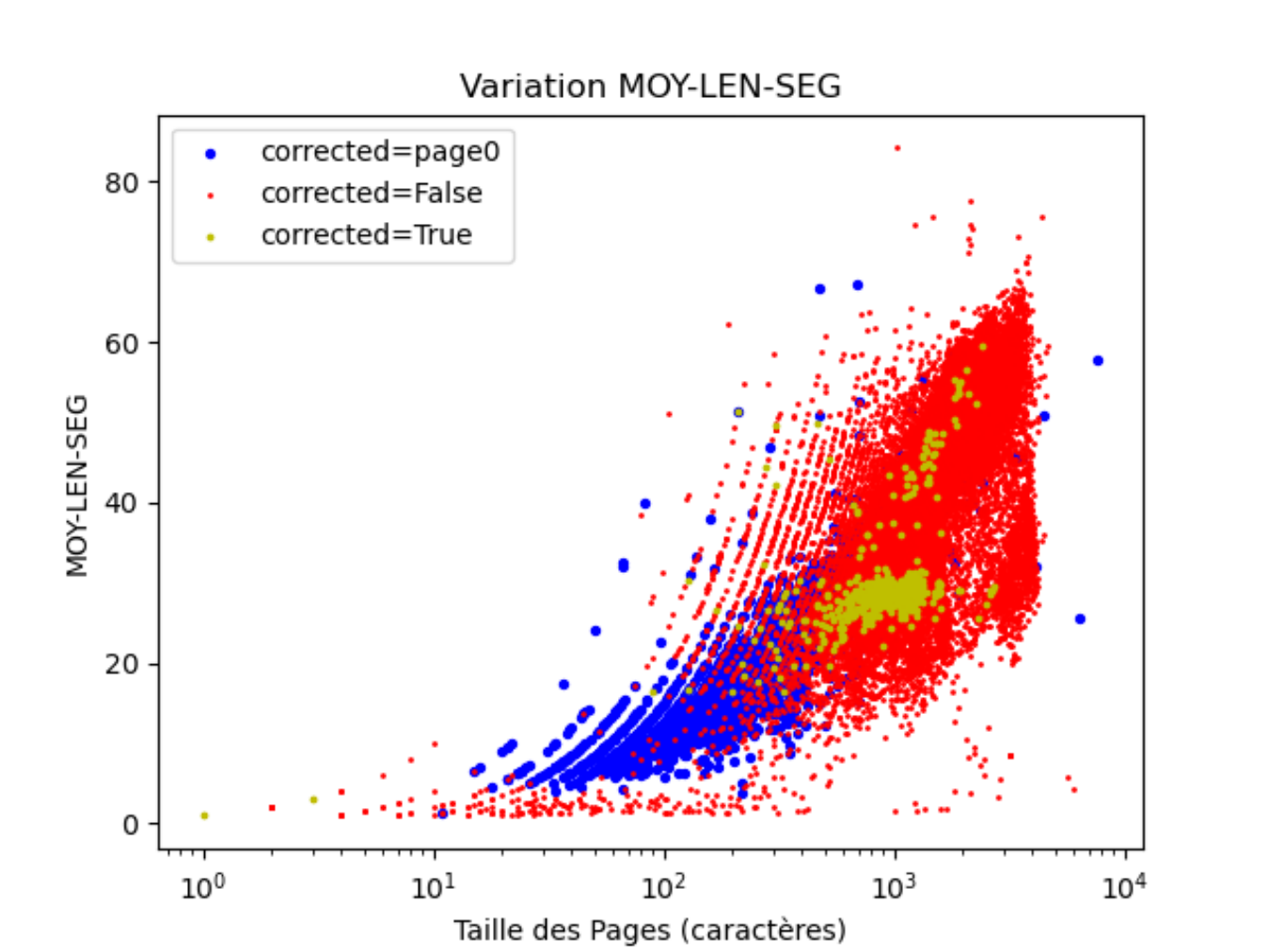


Figure 2 – Variation de la moyenne de la taille des segments

Comment mesurer l'étendue des dégats ? (II)

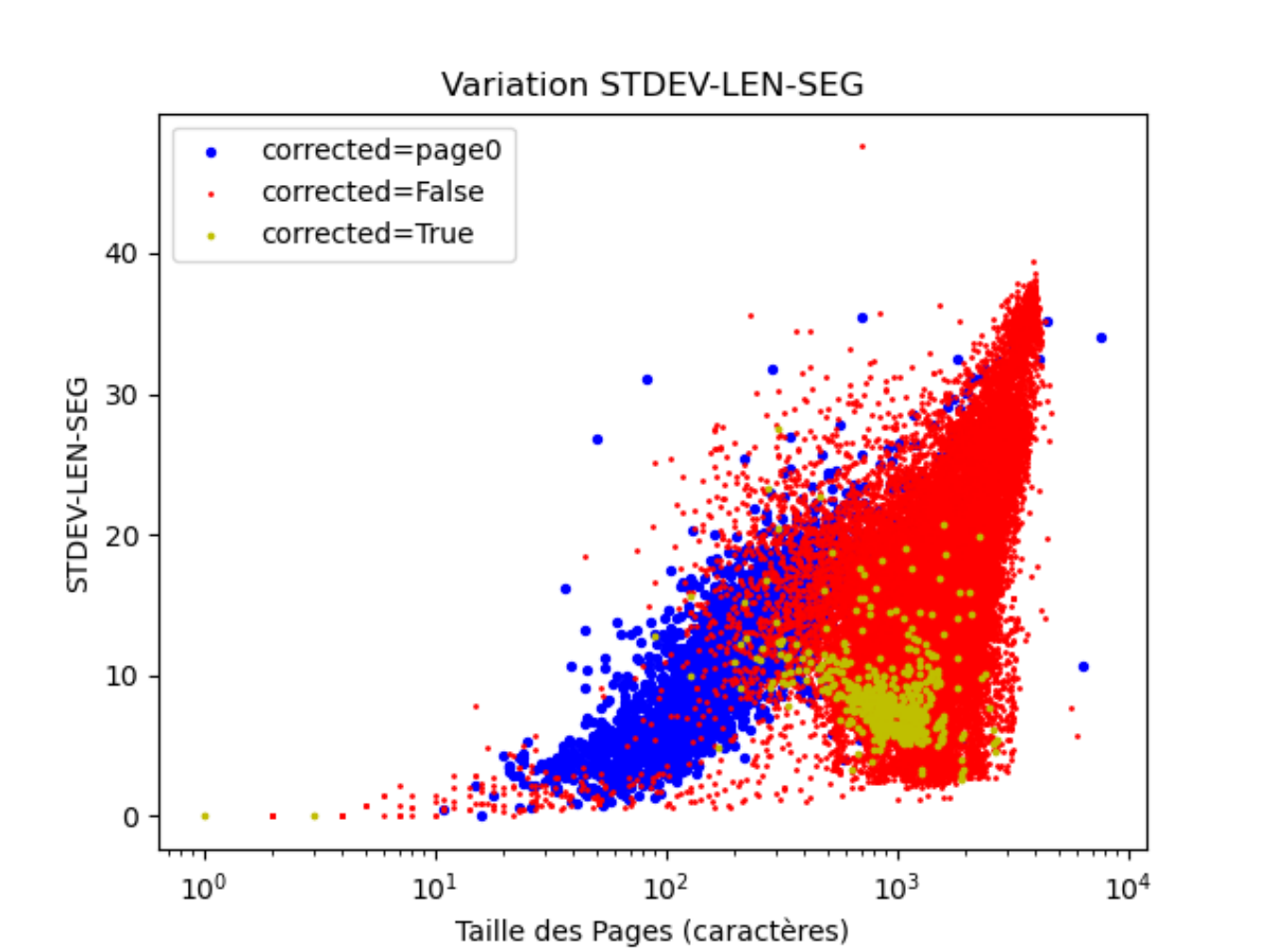


Figure 3 – Ecart-type de la taille des segments

Comment mesurer l'étendue des dégats ? (III)

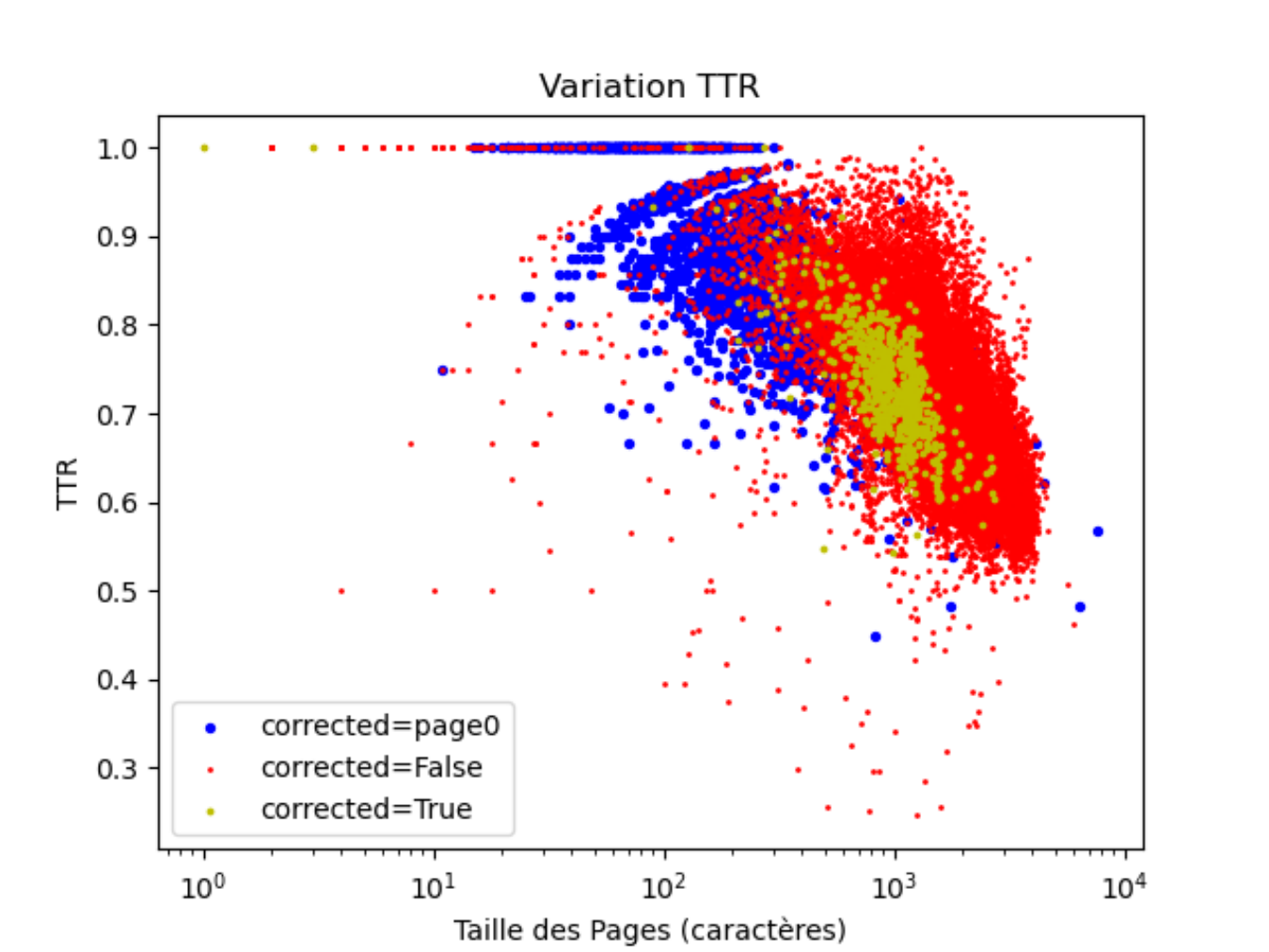


Figure 4 – Type/Token ratio (aka TTR)

Comment mesurer l'étendue des dégats ? (IV)

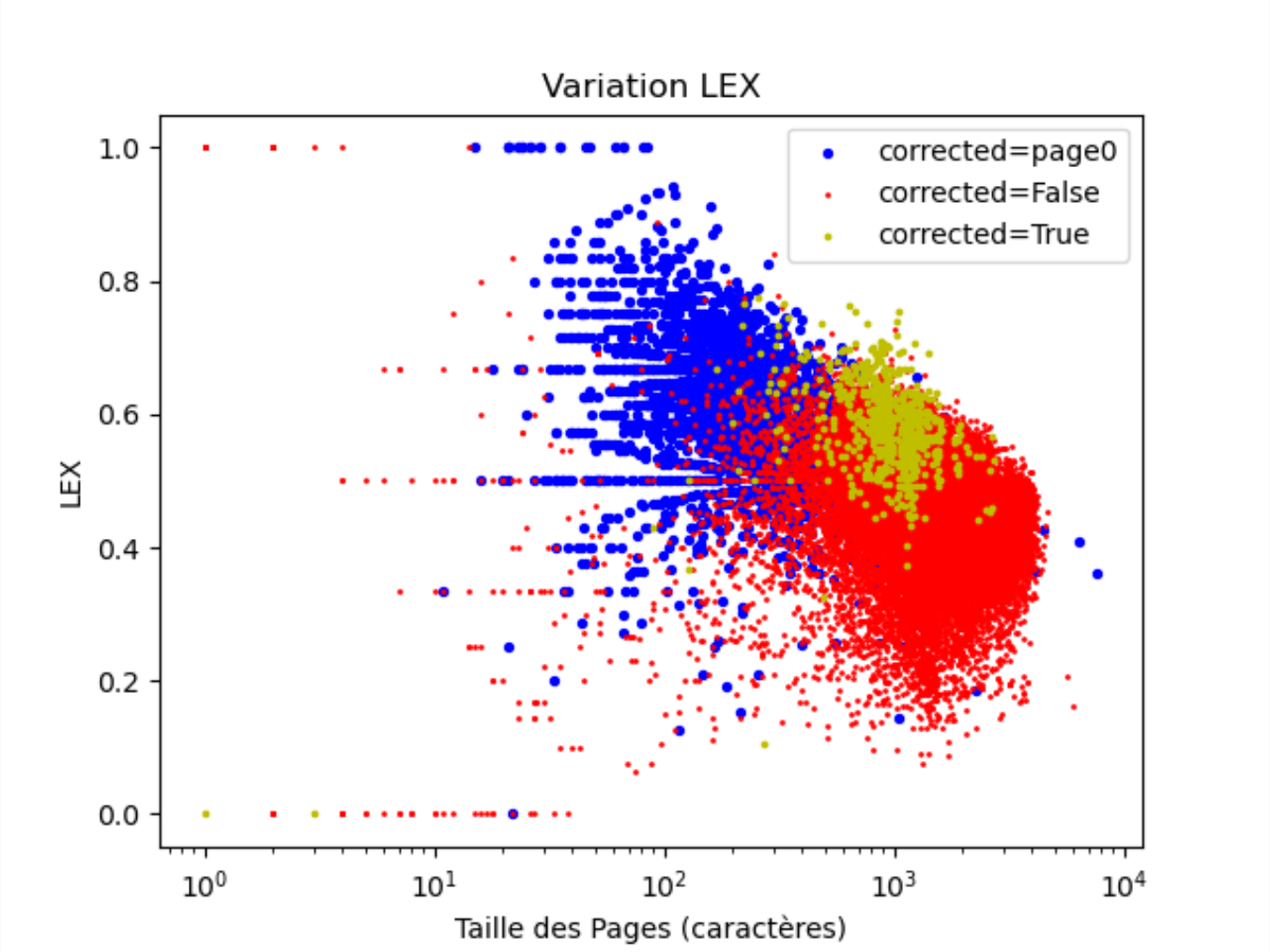


Figure 5 – Variation du Taux de Lexicalité (Ici : `LGERM`)

Comment mesurer l'étendue des dégats ? (V)

```
>>> import langid
>>> chaine = "Je m'appelle titi toto, j'adore la start-up nation when it comes to business et toi ? "
>>> for lg, score in langid.rank(chaine)[:10]:
...   print(lg, score)
...
en -270.39960765838623
la -285.9234924316406
br -297.9555459022522
fr -300.76536655426025
lb -310.3292818069458
af -314.26285791397095
it -314.32038736343384
es -314.55359411239624
oc -314.87990713119507
sw -314.98526668548584
```

Figure 6 – Utiliser le score de `langid` ?

Comment mesurer l'étendue des dégats ? (V)

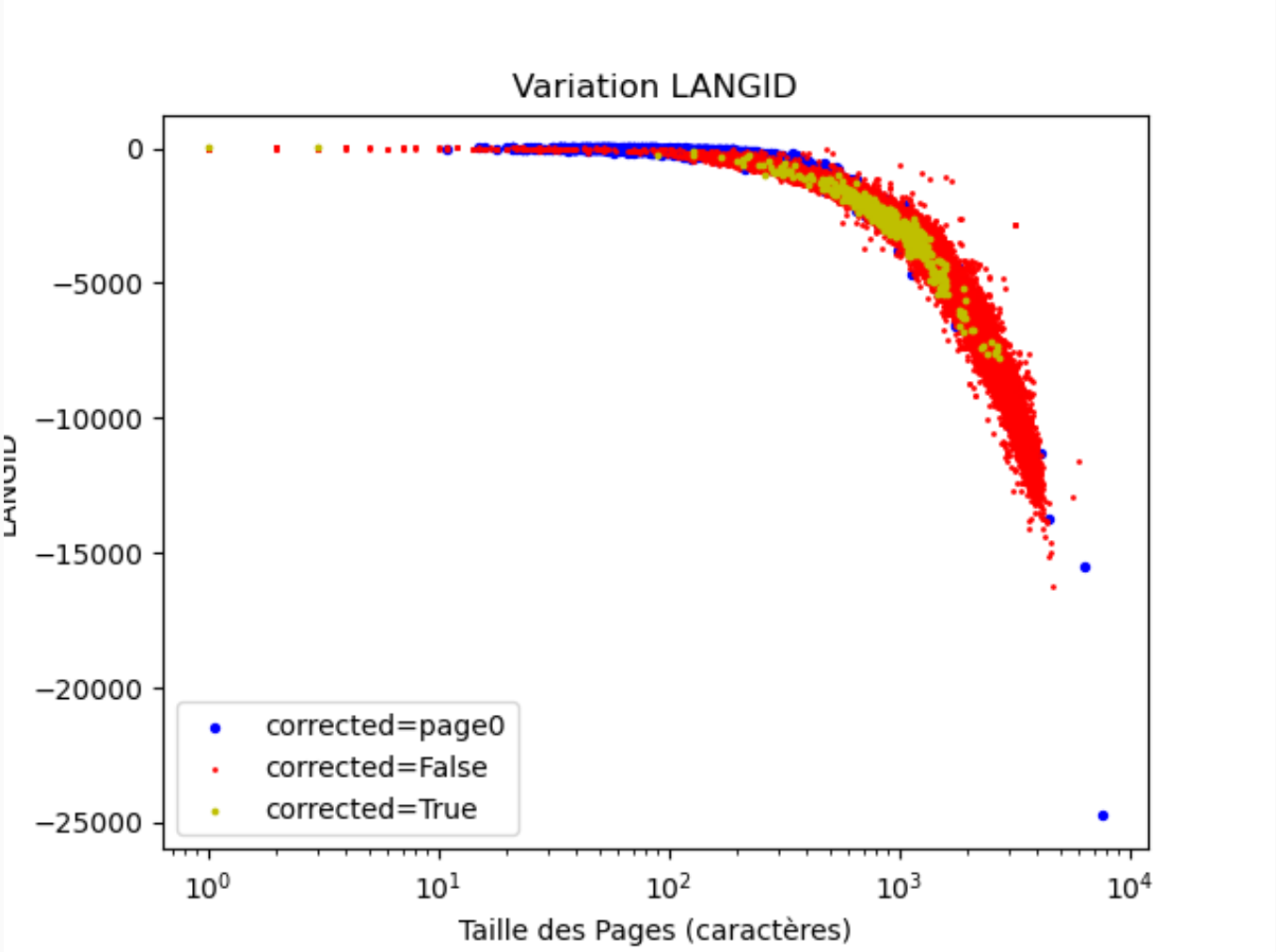


Figure 7 – Perturbation de l'identification de langue (Ici : `langid`)

Comment exploiter ces features ?

```
[11572, 'words 1SVC']
              precision    recall  f1-score   support

       False       0.94      1.00      0.97     11134
        True       0.95      0.40      0.56      1165

    accuracy                           0.94     12299
   macro avg       0.94      0.70      0.76     12299
weighted avg       0.94      0.94      0.93     12299

[11567, 'words 1ridge classifier']
              precision    recall  f1-score   support

       False       0.94      1.00      0.97     11134
        True       0.91      0.41      0.57      1165

    accuracy                           0.94     12299
   macro avg       0.93      0.70      0.77     12299
weighted avg       0.94      0.94      0.93     12299

[11547, 'words 1Perceptron']
              precision    recall  f1-score   support

       False       0.95      0.99      0.97     11134
        True       0.82      0.45      0.58      1165

    accuracy                           0.94     12299
   macro avg       0.88      0.72      0.78     12299
weighted avg       0.93      0.94      0.93     12299
```

Comment exploiter ces features ?

```
[11569, 'words 1SVC']
              precision    recall  f1-score   support

       False       0.94      1.00      0.97     11134
        True       0.43      0.01      0.03       228
       page0       0.96      0.49      0.65       937

    accuracy                           0.94     12299
   macro avg       0.78      0.50      0.55     12299
weighted avg       0.93      0.94      0.93     12299

[11545, 'words 1Perceptron']
              precision    recall  f1-score   support

       False       0.95      0.99      0.97     11134
        True       0.00      0.00      0.00       228
       page0       0.82      0.56      0.66       937

    accuracy                           0.94     12299
   macro avg       0.59      0.52      0.54     12299
weighted avg       0.92      0.94      0.93     12299

[11517, 'words 1ridge classifier']
              precision    recall  f1-score   support

       False       0.94      1.00      0.97     11134
        True       0.00      0.00      0.00       228
       page0       0.90      0.45      0.60       937

    accuracy                           0.94     12299
```

Comment exploiter ces features ?

```
[11630, 'words 1Perceptron_balanced']
              precision    recall  f1-score   support

       False       0.97      0.97      0.97     11134
        True       0.02      0.00      0.01       228
       page0       0.77      0.83      0.80       937

    accuracy                           0.95     12299
   macro avg       0.58      0.60      0.59     12299
weighted avg       0.93      0.95      0.94     12299

[11563, 'words 1SVC_bal']
              precision    recall  f1-score   support

       False       0.95      0.99      0.97     11134
        True       0.32      0.19      0.24       228
       page0       0.91      0.56      0.70       937

    accuracy                           0.94     12299
   macro avg       0.73      0.58      0.63     12299
weighted avg       0.93      0.94      0.93     12299
```

Figure 10 – Classification supervisée avec trois classes (class_weight="balanced")

Comment exploiter ces features ?

```
[11132, 'words 1SVC_bal']
              precision    recall  f1-score   support

       False       0.99      0.99      0.99     11191
        True       0.26      0.17      0.21       188

    accuracy                           0.98     11379
   macro avg       0.62      0.58      0.60     11379
weighted avg       0.97      0.98      0.98     11379

[10869, 'words 1Perceptron_balanced']
              precision    recall  f1-score   support

       False       0.98      0.97      0.98     11191
        True       0.02      0.04      0.03       188

    accuracy                           0.96     11379
   macro avg       0.50      0.51      0.50     11379
weighted avg       0.97      0.96      0.96     11379
```

Figure 11 – Classification supervisée avec deux classes, sans page0 (class_weight="balanced")

Ma conclusion ?

Ma conclusion ?

Pas encore de conclusions, il faut bosser . . .

Des idées ?