ATELIER OCR

Correction automatique des sorties d'OCR

Ljudmila PETKOVIĆ

25 novembre 2021

Introduction

Contexte

- Stage effectué à l'OBVIL (29.03.2021-30.07.2021)
- Sous l'encadrement du D^{r} Motasem Alrahabi, ing. de recherche
- Améliorer la qualité de l'OCR dans la TGB pour la REN
- Mesurer l'impact de la correction d'OCR sur la REN

1. Effectuer l'état de l'art sur la correction des sorties OCR
2. Créer une bibliographie thématique Zotero
3. Choisir un outil de la correction d'OCR le plus efficace
4. Tester `spacy` avant et après chaque correction

TGB [1]

- Grand volume de documents XML, océrisés et non corrigés
- Issus des collections Gallica de la BnF
- 58 287 auteurs
- 128 441 imprimés français (principalement du XIX^{e} s.)

1. Très Grande Bibliothèque — site web

TGB : chronologie

Dates :

- XVII^{e} : 24
- XVIII^{e} : 7294
- **XIX^{e} : 95479**
- XX^{e} : 54

TGB : sous-corpus

Thématiques (classification Dewey) :

- **Littérature (Belles-lettres) : 35710 documents**
- Histoire de la France (depuis 486) : 28885
- Droit : 23776
- Économie domestique. Vie à la maison : 19622
- Les arts : 5653
- Astronomie et sciences connexes : 4307
- Journalisme, édition. Journaux : 3824
- Religion : 2576
- Langues romanes. Français : 1861
- Philosophie et disciplines connexes : 1491

Pourquoi corriger la sortie OCR ?

- Prérequis pour effectuer les divers tâches en TAL (p. ex. REN)
  - P**a**ris `[LOC-B`[2] `|` `VILLE]` ; P**8**ris [?]
- Les erreurs d'OCR ont un impact sur la performance de ces tâches[3]
- « Pré-nettoyage » manuel : fastidieux et chronophage
- Automatiser le « débruitage » (angl. *denoising*, Rigaud *et al.*, 2019)
- Erreurs systématiques (même caractère : même erreur)
  - Prédiction via la sortie de l'échantillon (Mokhtar et al., 2018)

2. Selon le schéma d'annotation BIO — *Beginning-Inside-Out* (Dupont, 2017)
3. Mesure F_1 : 0.76 pour les entités géopolitiques (van Strien *et al.*, 2020).

Problématique

- La correction de sortie OCR est une tâche non triviale
- Systèmes performants (CER : 0.01%, Reul *et al.*, 2018 ; 2019) :
  - Kraken, Transkribus, Tesseract, Calamari, ABBYY...
- Les erreurs d'OCR sont inévitables (et souvent assez nombreuses) :
  - Mauvaise qualité des images-sources
  - Mise en page complexe
  - Polices historiques (ligatures, variantes orthographiques)...
  - ...

Typologie des erreurs [4]

Faux positifs

- Erreurs humaines incorrectement corrigées par le correcteur ; surcorrections (angl. *real-word errors*)
- Typiquement dues à la grande largeur des dictionnaires
- *veery* (la grive fauve) > *very*

Figure 1 – La grive fauve, source : Wikipedia.

4. Selon Edwards (2016).

Typologie des erreurs

Faux négatifs

- Mots détectés par le correcteur comme erronés (angl. *non-word errors*) alors qu'ils ne le sont pas du point de vue de l'humain
- Dans le cas d'un dictionnaire trop étroit
- *IQN*
  - International Quality Network (scientific research)
  - iSCSI Qualified Name
  - Integrating Quality and Novelty (peer-to-peer search engine)
  - Industrial Química del Nalón SA (Spanish chemical company)
  - ...

Typologie des erreurs [5]

Erreurs grammaticales

- Infinitif au lieu du participe passé : « J'ai _____une pomme »

Erreurs de sens

- Homonymes (*chat* à la place de *chah* dans « le chah d'Iran »)
- Typographie/coquilles (*poison* et *poisson*)
- Mauvaise délimitation entre deux mots (*puis que* et *puisque*)

Erreurs de syntaxe

- Répétition d'un mot par inadvertance (doublon) : *de* dans « erreurs *de de* syntaxe »
- Omission d'un mot (bourdon) : *de* omis de la phrase « erreurs syntaxe »

5. Selon Dumas (2017).

Typologie des erreurs [6]

- Cognitives : *Levenstain* > *Leven**sh**te**i**n*
- D'OCR : ***IN**fo**h**ma**T**ion* > *information*
- Diacritiques (*Cedric* > *C**é**dric*).
- Du slang : *dsl* > *désolé·e* (Bhashkar, 2019)

6. Selon Hládek *et al.* (2020).

Bibliothèque Zotero

État de l'art

- différentes méthodes d'OCR / d'HTR, de la correction d'orthographe, avec les aperçus (angl. *surveys*) de la correction d'OCR (ex : Hládek *et al.*, 2020) ;
- les études abordant de manière générale les mécanismes de correction d'orthographe (p. ex. Kukich, 1992) ;
- les études de cas sur la correction d'OCR (p. ex. la correction les fautes d'orthographe des mots réels en rétablissant la cohésion lexicale de Hirst & Budanitsky (2003)) ;
- l'impact de l'OCR sur les tâches numériques (ex : Traub *et al.*, 2015).

Flux de travail

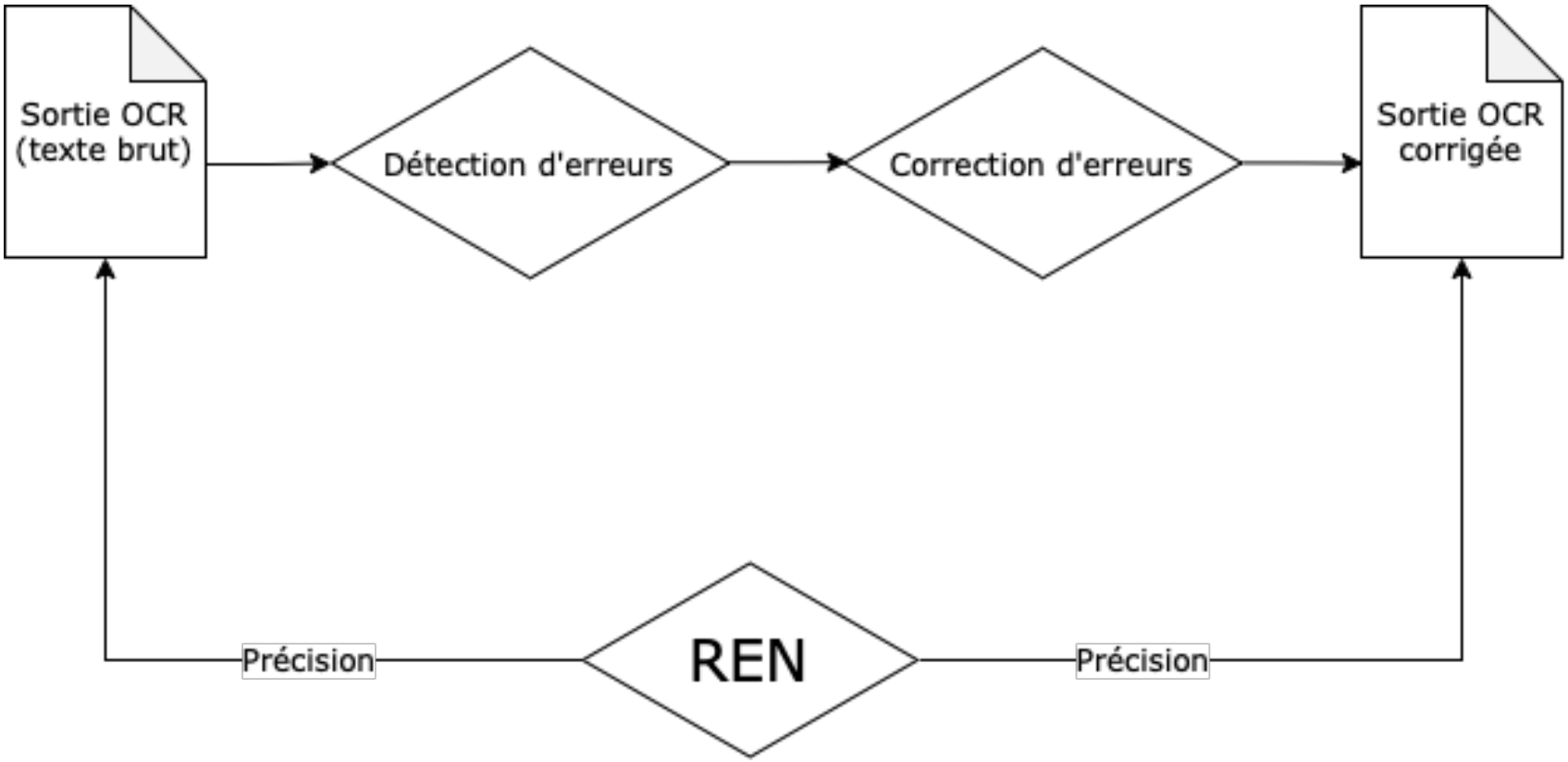

Figure 2 – Diagramme de flux de la correction post-OCR.

État de l'art

Approches de correction de sortie OCR

- Manuelle (correction participative)
- Lexicale
  - Dictionnaires
  - Distance d'édition
  - Règles
  - Syntaxique
- Probabiliste (statistique, modèle de langage, n-grammes)
- Apprentissage machine / profond
- Hybride

Manuelle

- Relecture et rectification des erreurs par l'intervention humaine
- Correction participative ou collaborative (angl. *crowdsourcing*)
- Produire un corpus étalon-or (angl. *gold standard corpus*)
- Bibliothèques numériques (p. ex. la BnF)
- Haute précision, voire la qualité éditoriale
- Augmentation du nombre de docs récupérés
- Chronophage
- Manque de docs-source (référence aux corpus océrisés)
- Efficacité et expertise des bénévoles variables
- Distinguer des caractères similaires (ex. *i*, *l*, *I*, *1*, *0*, *O*, *o*, *e*, *c*)

Dictionnaires

- Parcourir un lexique de recherche pour chaque mot du texte
- Vérifier si un mot y est répertorié, autrement dit si son orthographe correspond à celle du mot recherche dans le dictionnaire
- Si non, il est considéré comme un mot erroné à corriger
- Générer une liste de candidats pour corriger les mots
- *bnojour* > *séjour*, *abat-jour*, *bonjour* (candidats de correction)
- Correction interactive possible (`aspell`)

Aspell

Usage basique :

1. Télécharger le logiciel
2. Décompresser le package `aspell-0.60.8.tar`
3. Naviguer vers le dossier en ligne de commande
4. `./configure  make`
5. `make install`
6. `aspell -mode=sgml`[7] `-c test.xml`

7. Filtre pour traiter les documents génériques SGML/XML.

Aspell

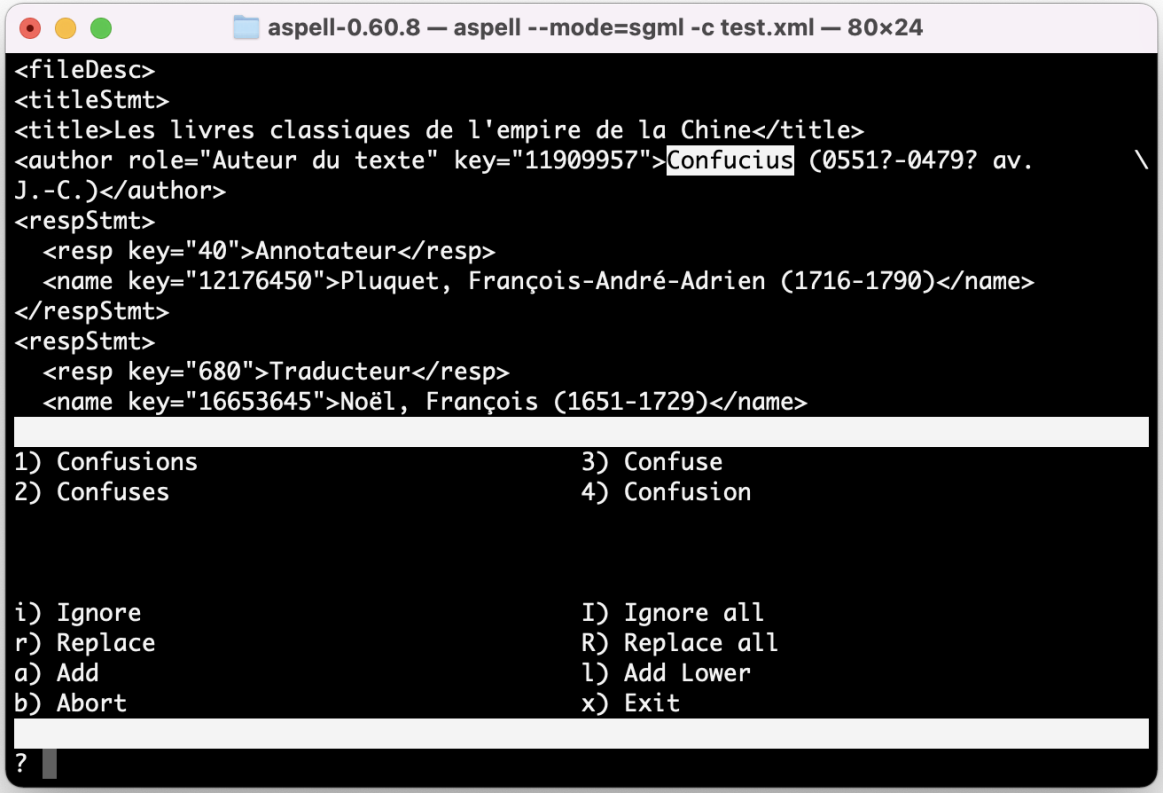

Figure 3 – Correction d'orthographe interactive par `aspell`.

Distance d'édition

Définition

Nombre minimum d'opérations d'édition nécessaires pour réécrire un mot *w* en mot *w'* (Tantini *et al.*, 2011) ∼ *CER*

Opération	Mot incorrect	Mot correct
Insertion	miot	mot
Suppression	mt	mot
Substitution	mof	mot
Transposition (inversion)	mto	mot

Table 1 – Opérations effectuées au sein d'une chaîne de traitement illustrant la distance d'édition Damerau-Levenshtein.

Ex. : `pyspellchecker`, `pyenchant`...

Règles

- PoCoTo (Vobl *et al.*, 2014)
- Drobac, 2017 ; 2020
- Volk *et al.*, 2011 :
  - similarités graphémiques : ***K**edaktion* > ***R**edaktion*
- Wisniewski *et al.*, 2010 :
  - Récupérer automatiquement les archives de l'encyclopédie collaborative Wikipédia contenant les corrections orthographiques
  - Améliorer la liste des candidats suggérés pour la correction des mots erronés au sein du logiciel `hunspell`

Règles

LanguageTool (Naber, 2003)

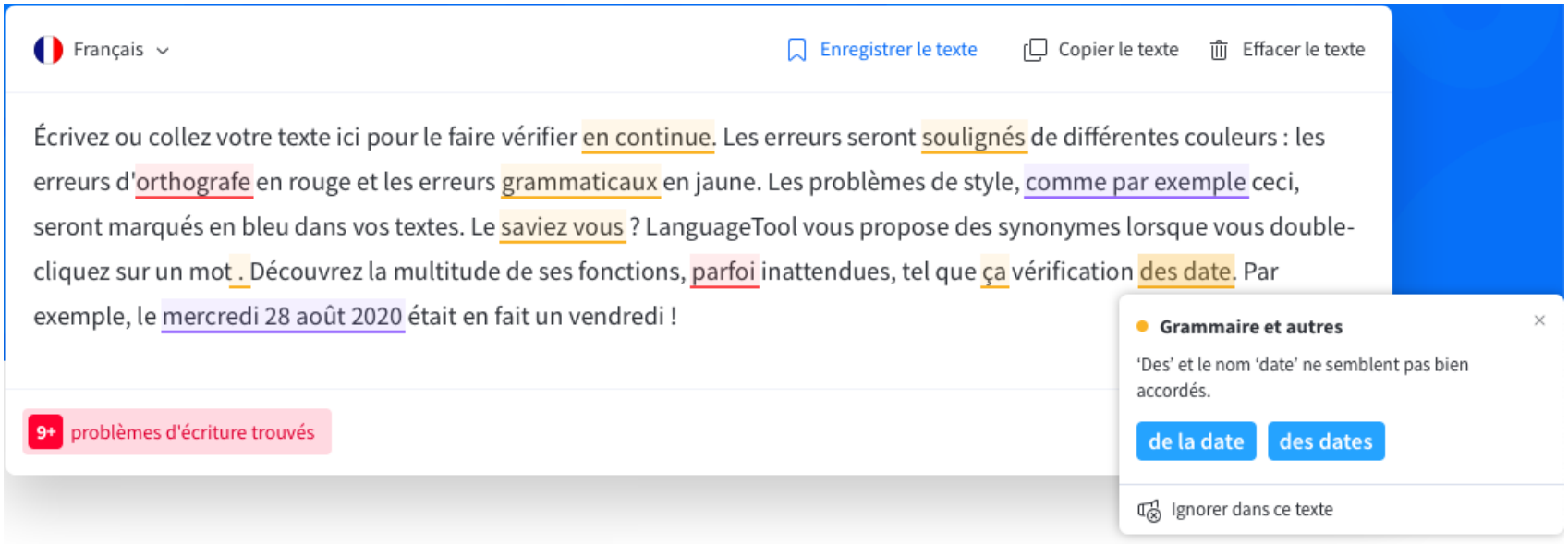

Figure 4 – Correction d'orthographe interactive en tenant compte des règles linguistiques.

Règles

- Liste des erreurs / corrections annotées (tableur .csv)
  - Annotations effectuées par les étudiants / stagiaires OBVIL

Syntaxique

- Décomposition et désambiguïsation d'une phrase
- Correction des erreurs dépendantes du contexte
- Lexique, grammaire, processus d'analyse
- *Le marin est courageuse*

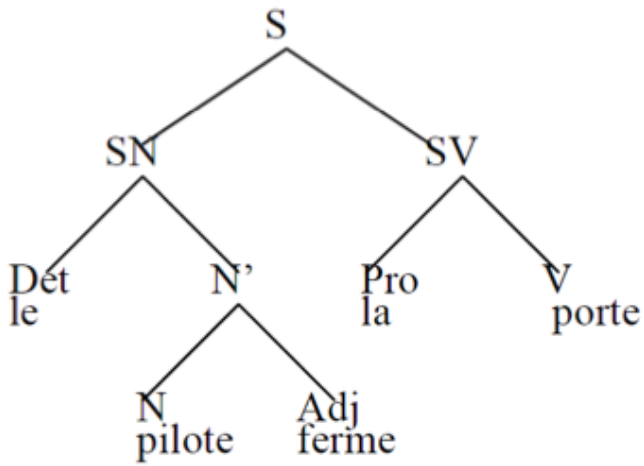

Figure 5 – Le pilote qui est ferme porte quelque chose.

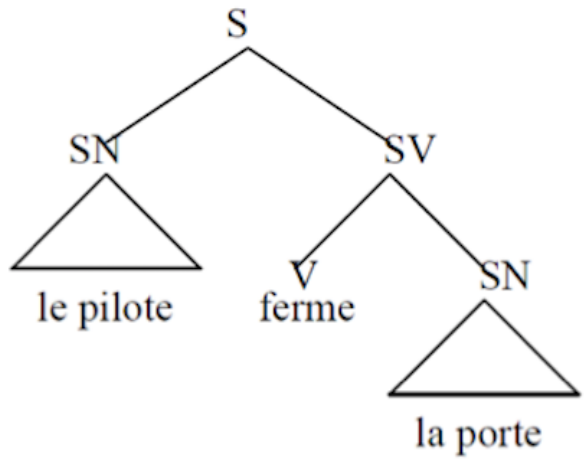

Figure 6 – Le pilote effectue la fermeture de la porte.

Probabiliste

- Modélisation statistique du langage et les n-grammes de mots
- Calculer la probabilité d'apparition d'une séquence de mots
  - Bigramme : *le chat* — la probabilité d'un nouveau mot (*chat*) dépend du mot précédent (*le*)
  - Trigramme : *le chat miaule* — la probabilité d'un nouveau mot (*miaoule*) dépend des probabilités des deux mots précédents (*le chat*) ...
  - *ce**l**te dénégation de l'appelant* : ***cette*** ou *celte*
- Matrice de fréquences d'apparition de n-grammes calculés à partir d'un corpus
- Cascades et CRF (Mubarak & Darwish, 2014)
- Algorithme de Google *Did you mean* (Bassil & Alwani, 2012)

Probabiliste

- Correcteur orthographique contextuel pour la langue assamaise (Choudhury *et al.*, 2019)
- Optimisation pour la vitesse / la consommation de mémoire, grâce au *filtre de Bloom* et au *hachage parfait*
- Faible gestion des dépendances à longue distance
  - Mots qui dépassent la portée d'un modèle de n-gramme
  - Quel livre Pierre doit-il lire ?
  - Solution : modèles de langage basés sur l'analyse syntaxique

Apprentissage

- seq2seq [8] (Mokhtar *et al.*, 2018, Salimzadeh, 2019)...
  - au niveau des caractères + couches LSTM
- BERT [9] (Tan *et al.*, 2020)
  - modèle de représentation linguistique
  - aperçu
- BERT + NMT [10] (Nguyen *et al.*, 2020)

Aperçu de l'état de l'art : Nguyen *et al.*, 2021

8. Angl. *sequence-to-sequence*.
9. Angl. *Bidirectional Encoder Representations from Transformers* (Devlin *et al.*, 2018).
10. Angl. *Neural Machine Translation*.

Apprentissage

- Réglage fin : utiliser les poids d'un réseau de neurones déjà entraîné et l'utiliser comme une base pour un nouveau modèle en cours d'entraînement sur les données du même domaine
- Désambiguïsation du contexte
  - I will call my siter. [*sister*]
  - Due to bad weather, we had to move to a different siter. [*site*]
- Nécessitent un énorme ensemble de données d'entraînement
- Surapprentissage : incapacité de généralisation
  - Si le modèle est exclusivement entraîné sur un type de données

Hybride

- SxPipe (Gábor & Sagot, 2014)
  - algorithmes de classification statistique + connaissances linguistiques (cascade de traitements superficiel [11])
- Modèle de canal bruité (approche statistique) + classifieurs Winnow (apprentissage machine) (Daðason, 2012)

11. Angl. *shallow processing*.

Résumé des approches de correction d'OCR

Approche	Avantages	Désavantages
Manuelle	• Précision d'étalon-or • Correction collaborative	• Chronophage • Indisponibilité des documents de référence • Efficacité et expertise des correcteurs variables
Lexicale	• Création facile du dico • Extensibilité des dicos	• Chronophage si le dico et/ou le texte est grand • Incomplétude des dicos • Indépendance du contexte
Probabiliste	• Dépendance du contexte	• Gestion des dépendances à longue distance
Apprentissage	• Dépendance du contexte • Apprentissage par transfert • Réglage fin	• Nécessité d'un grand corpus d'entraînement • Surapprentissage • Données d'apprentissage clairsemées

Méthodologie

Pré-traitements

Pour éviter les surcorrections (p. ex. `Confucius` > `Confucrus`)

1. Transformation du fichier TEI-XML au fichier .txt
   - Enlever les balises XML avec la librairie (`lxml`) ;
   - sinon `<div>` > *diva*
   - Ne corriger que le texte brut
2. Gestion des caractères spéciaux
   - Effacer les caractères spéciaux : •, %, *, #, +, $... générés par l'OCR, mais sans valeur sémantique
   - Regex `«.\n`
     - `un fpectacle. " Ah 1 s'écria Confu«`
       `cius, je n'avois pas vu jufqu'ici`
   - Minusculisation : *Chine* > *chine*
   - Réduction des espaces multiples
   - Guillemets : ’ > ‘ ; sinon `l'empire` > `laempire`
   - `jeuneffe,` > `jeuneffe` ; sinon, pas de correction
   - Tokeniseur standard de Python ; sinon `l'empire` > `l' empire`

Pré-traitements

- Omission des parties du texte dans les balises imbriquées
- Récupération de ces parties avec l'attribut `.tail`

```
<p rend="small">antiquité ; on l'attribue en grande
partie à Fo - hi : c'eft un ouvrage
qui, par le moyen des <hi rend="i">emblèmes</hi>,
explique ou repréfente la doétrine
..les effets de cette vertu ( i).</p><p
rend="small">(i) Notice de l'Y-king, par M.
Vifdeîau,[...]
```

Approche par dictionnaire

- `pyspellchecker`
- `pyenchant` (stage de Nicolas Hiebel — dépôt GitHub)
  - Corrections des mots ignorés par `pyspellchecker`
  - P. ex. *néce**fl**î**t**é* > *nécessité* ; *defendre* > *d**é**fendre*…
  - …mais *confucius* > *confu**si**o**n**s*

pyspellchecker

Erreur	Correction	Fréquence
dirfufion	diffusion	1
fubtilité	subtilité	1
confufion	confusion	1
fophe	force	1
cnfcignement	cnfcignement	1
illupcres	illustres	1
s'étoit	setait	1
falle	fille	1
doéleurs	douleurs	1
rétablifsent	rétablissent	1

Table 2 – Extrait des corrections automatiques des tokens mal orthographiés par `pyspellchecker`.

Approche par apprentissage

`jamspell`

- Correcteur orthographique basé sur l'apprentissage machine
- Entraîné sur les trigrammes multilingues pour corriger et sélectionner le candidat avec le score le plus élevé
- Plus rapide que `pyspellchecker` (quelques secondes vs. 15 mins pour un fichier 100-200 Ko)
- Démonstration Google Colab

jamspell

Erreur	Correction	Fréquence
renfermoient	renfermoient	1
refpecl	respect	1
pouvoit	pouvoir	1
puifsance	puissance	1
0"uvernement	0"gouvernement	1
doétrine	doctrine ou doétrine	1
tchun-tfiou	chun-tfiou	1
foiblefse	faiblesse	1
priacipautéde	pricipautéde	1
confidéroit	considérait	1
cmpire	empire	1

Table 3 – Extrait des corrections automatiques des tokens mal orthographiés par `jamspell`.

Futures recherches

Futures recherches

- OCR-NE (dépôt GitHub)
- Entraîner le modèle de français avec `neuspell`

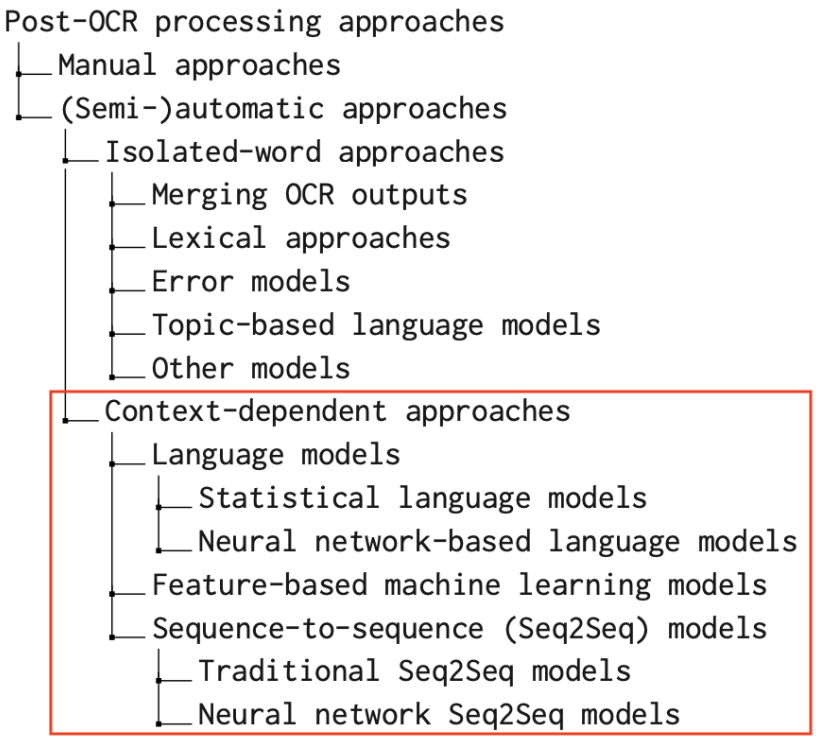

Figure 7 – La taxinomie des approches de correction de sortie OCR (Nguyen *et al.*, 2021).