**Journée d’étude métadictionaire :**

**La reconnaissance automatique de caractère entre approche industrielle et spécialisée[[1]](#footnote-1).**

**Anaïs :**

En introduction, il me semble que nous pourrions commencer par replacer les « enjeux numériques » d’un tel projet : mettre à disposition des dictionnaires numérisés, open access, faciliter la recherche, etc.

Ensuite, développer les moyens techniques mis en œuvre, les difficultés rencontrées et les solutions employées.

- pour le traitement spécifique des vedettes grecques, nous avons le Thevenin, le De Gorris et l’index du De Gorris (1601-1658)

- pour le traitement des vedettes latines suivies d’étymologies grecques le Castelli et les deux James (1743-1746)

- il y a également les dictionnaires plus récents qui ont été pris en charge par Frédéric ainsi que le développement de l’interface

Je vais t’exposer rapidement ici ce que j’ai mis en œuvre de mon côté.

1.       Un script python de traitement automatique des images en sortie de scanner HD (suppression du bruit, ajustement des niveaux et des contrastes, amélioration de la netteté à l’échelle d’un pixel)

2.       Un script python de détection automatique du texte sur une image de page afin de procéder à un découpage au plus proche du contenu (très utile pour les grands formats)

3.       Une comparaison de l’efficacité Transkribus/ABBYY sur des papiers abîmés —> recherche d’un modèle hybride (manuscrit/imprimé) et multilingue (grec/latin)

4.       Mise en place de stratégies de segmentation des pages suivant les documents (modèle simple pour le Gorris et son index car les colonnes ont pu être séparées au préalable grâce au script ; modèle semi-automatisé pour le Thevenin car colonnes peu définies)

5.       Tests de différents modèles HTR

a.       Utilisation d’un seul et même modèle, car entrainé sur des données d’une période similaire : le Noscemus GM5 développé dans le cadre d’un projet européen par Stefan Strahammer.

6.       Sortie DOCX pour le De Gorris et le Thevenin. J’aurai voulu isoler les vedettes grâce à leur typographie finale, mais ce n’était pas assez systématique pour être efficace…

a.       regex et stylage LO optimisé

b.       passage dans Odette (développé par Frédéric)

c.       application d’un script XSL afin de générer les différents attributs et normaliser en TEI

7.       Sortie TXT pour l’index du De Gorris.

a.        application d’un script python csv —> XML pour generer automatiquement le glossaire

**Cahal**

Récupération du contenu de trois dictionnaires avec Abby pour avoir en sortie un fichier XML

Préambules :

1- Étude de la qualité du scan, peut-on améliorer ? Y a-t-il une meilleure version (exemple du James Fr dans Gallica dont la qualité permet un gain de reconnaissance important) ?

2- Effectuer une première génération avec les gabarits « stock » d’Abby, constat, plus le texte est ancien et donc la technique de l’édition s’éloigne de la notre moins le résultat est bon (exemple du Castelli notamment sur la segmentation des mots).

Apprentissage :

1- Élément récupérable, l’italique donc une partie de la taxinomie du dictionnaire en plus du texte.

2- Premier apprentissage sur les caractères latins, amélioration de la reconnaissance et adaptation à la spécificité du texte, le cas des S longs.

3- Les caractères grecs prisent en compte des ligatures spécifique au texte Cf. N. Oikonomides, Manuel, Abbreviations in Greek Inscriptions : Papyri, Manuscripts, and Early Printed Books.

a-Amélioré les caractères normaux et faire apprendre les ligatures, effort général puisqu’une ligature inconnue ne pourra pas être récupérée. Exemple des diverses ligatures qui en sortie donne du texte Unicode.

b-Autres problèmes le placement des diacritiques divergent des pratiques éditorial de nos jours (exemple d’un accent non existant en Unicode, mais utilisé à l’époque)

Export du texte :

Choisir le bon format d’export pour obtenir un texte continu conservant l’italique, mais sans autre élément pouvant compliquer l’XML qu’on convertira.

Envoie du texte à Frédéric parti sur l’édition systématique du XML par des commandes de *search and replace* à voir avec lui.

**Commun :**

Comparaison des méthodes sur un segment de texte, comparer les gains et les pertes, le bruit et autre point divergent ainsi que les investissements de temps et de ressources (besoin d’ordinateur puissant pour un OCR massif, etc.).

Analyse des données statistiques d’Abby sur les environ 35 0000 pages que j’ai traitées (Gallien latin, Castelli, James FR et ENG).

1. Titre indicatif. [↑](#footnote-ref-1)