Les outils numériques au service de la philologie

Sophie Robert-Hayek

Laboratoire Écritures, Maison des Sciences de l'Homme, Université de Lorraine

17 Septembre 2024

Séminaire Écritures

Plan de la présentation

Qu'est-ce que les humanités numériques ?

Définition

Les humanités numériques

Les **humanités numériques** sont :

- un domaine interdisciplinaire;
- qui associe la recherche en sciences humaines traditionnelles;
- avec des **outils issus des sciences numériques**;
- pour tenter de donner des réponses à des problématiques issues des sciences humaines.

Définition

Les humanités numériques

Les **humanités numériques** sont :

- un domaine interdisciplinaire;
- qui associe la recherche en sciences humaines traditionnelles;
- avec des **outils issus des sciences numériques**;
- pour tenter de donner des réponses à des problématiques issues des sciences humaines.

Les humanités numériques rassemblent des experts issus d'un large éventail de disciplines :

- **sciences humaines**
- **mathématiciens**
- **informaticiens**

pour tenter d'apporter de nouvelles réponses et de nouveaux angles aux problèmes existants.

Qu'est-ce que les humanités numériques ?

Les avancées récentes en informatique offrent des opportunités **sans précédent** pour

- **Générer**

Qu'est-ce que les humanités numériques ?

Les avancées récentes en informatique offrent des opportunités **sans précédent** pour

- **Générer**
- **Explorer**

Qu'est-ce que les humanités numériques ?

Les avancées récentes en informatique offrent des opportunités **sans précédent** pour

- **Générer**
- **Explorer**
- **Interpréter**

des données.

L'intégration de l'informatique avec les disciplines des sciences humaines promet **de nouvelles perspectives pour la recherche, l'analyse et la compréhension des données existantes**.

Qu'est-ce que les humanités numériques ?

La convergence de l'informatique et des sciences humaines peut permettre :

Qu'est-ce que les humanités numériques ?

La convergence de l'informatique et des sciences humaines peut permettre :

- De marquer une nouvelle ère de **collaboration interdisciplinaire** pour aborder les questions de recherche sous différents angles;

Qu'est-ce que les humanités numériques ?

La convergence de l'informatique et des sciences humaines peut permettre :

- De marquer une nouvelle ère de **collaboration interdisciplinaire** pour aborder les questions de recherche sous différents angles;
- **D'élargir les perspectives de recherche** grâce à des initiatives de données ouvertes et des plateformes collaboratives;
- De **proposer de nouvelles méthodes quantitatives** pour répondre à des **questions existantes en sciences humaines**.

Humanités numériques

Les projets en humanités numériques peuvent être grossièrement divisés en trois grandes catégories :

Humanités numériques

Les projets en humanités numériques peuvent être grossièrement divisés en trois grandes catégories :

- **Application de l'intelligence artificielle/mathématiques** : application de modèles mathématiques pour mieux comprendre les données des sciences humaines ;

Humanités numériques

Les projets en humanités numériques peuvent être grossièrement divisés en trois grandes catégories :

- **Application de l'intelligence artificielle/mathématiques** : application de modèles mathématiques pour mieux comprendre les données des sciences humaines ;
- **Application de l'ingénierie des données** :
  - structurer les données à partir de données physiques/non structurées ;
  - définir de nouvelles normes de données au sein de la communauté de recherche.

Humanités numériques

Les projets en humanités numériques peuvent être grossièrement divisés en trois grandes catégories :

- **Application de l'intelligence artificielle/mathématiques** : application de modèles mathématiques pour mieux comprendre les données des sciences humaines ;
- **Application de l'ingénierie des données** :
  - structurer les données à partir de données physiques/non structurées ;
  - définir de nouvelles normes de données au sein de la communauté de recherche.
- **Application de l'ingénierie logicielle** :
  - développer des logiciels pour faciliter l'accès et la manipulation des données ;
  - concevoir de nouvelles façons d'interagir avec les données pour en tirer des connaissances.

La philologie computationnelle à Écriture

Les humanités numériques à Écriture

La philologie est dîte **computationnelle** quand elle est réalisée de manière **automatique** ou **semi-automatique** à l'aide des sciences numériques.

Les humanités numériques à Écriture

La philologie est dîte **computationnelle** quand elle est réalisée de manière **automatique** ou **semi-automatique** à l'aide des sciences numériques.

Comme le reste des humanités numériques, elle peut se décliner sous la forme:

- D'**ingénierie des données** (création de nouvelles bases de données...);
- D'**ingénierie logiciel** (développement de logiciels permettant d'aider le philologue dans sa tâche);
- D'**utilisation d'algorithmes d'informatique et de mathématiques appliquées** pour répondre à des problématiques d'humanités.

Les humanités numériques à Écriture

Trois projets en philologie computationnelle en cours:

- **Le projet SCRIBES** : développement d'une application Web pour la construction d'éditions critiques;

Les humanités numériques à Écriture

Trois projets en philologie computationnelle en cours:

- **Le projet SCRIBES** : développement d'une application Web pour la construction d'éditions critiques;
- **Les projets BENTO (LUE) / SHERBET (ANR)** : application de la stemmatologie computationelle au cas de la transmission du texte de Ben Sira / de Qumrân.

Le projet SCRIBES

Le projet SCRIBES a pour visée de **proposer une application Web collaborative** permettant **d'éditer facilement** une édition critique.

Permet de:

- **Transcrire un texte** à partir de **photos de manuscrits**;
- **Visualiser les résultats de la transcription**:
  - En diplomatique;
  - En collation;
  - Sous la forme d'un stemma.

Fonction d'édition

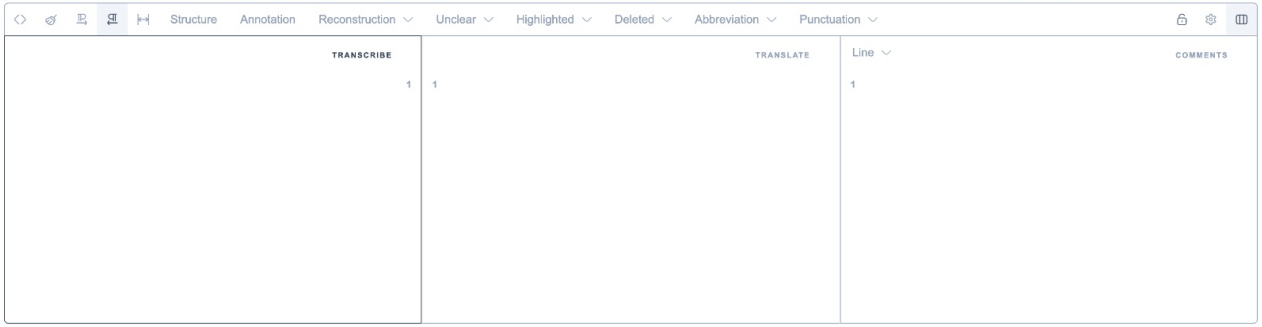
SCRIBES permet pour **une tradition textuelle donnée**:

- De téléverser une image de manuscrit;
- De réaliser la transcription, la traduction et la prise de note sur le texte;
- D'exporter le texte au **format XML**, de manière transparente.

Fonction d'édition

Fonction d'édition

Composant de transcription (en cours de développement):



Fonction de visualisation

Une fois la transcription réalisée, il est possible de visualiser les résultats:

- Diplomatique;
- Collation;
- Stemma.

Le projet SHERBET

SHERBET (Stemmatology for the HEbRew BiblE Transmission)

Financement de 4 ans (subvention ANR française) pour reconstruire les liens généalogiques des manuscrits de Qumran et de la Genizah du Caire en utilisant des **outils computationnels**.

Consortium de laboratoires de philologie (Écritures), de laboratoire d'informatique (LORIA) et de laboratoires de mathématiques appliquées (IECL, LJK).

Institut ÉLIE CARTAN

ÉCRITURES EA 3943

LABORATOIRE JEAN KUNTZMANN MATHÉMATIQUES APPLIQUÉES - INFORMATIQUE

Loria Laboratoire lorrain de recherche en informatique et ses applications

Qu'est-ce que la stemmatologie ?

Stemmatologie

La stemmatologie est la science qui vise à reconstruire l'arbre généalogique (stemma) des différents manuscrits d'un texte donné :

- Pour reconstruire un archétype (perspective reconstructionniste);
- Pour comprendre la transmission d'un texte à travers les siècles.

Phylogénie computationnelle

Phylogénie

La **phylogénie** consiste à construire un diagramme en arborescence pour montrer les relations entre les espèces biologiques ou autres entités. La biologie moderne utilise **les similitudes de leur séquence ADN**.

Phylogénie computationnelle

Phylogénie

La **phylogénie** consiste à construire un diagramme en arborescence pour montrer les relations entre les espèces biologiques ou autres entités. La biologie moderne utilise **les similitudes de leur séquence ADN**.

L'objectif principal de la phylogénie est **de reconstruire l'arbre généalogique le plus probable étant donné une liste de séquences ADN**.

...si nous considérons les manuscrits comme une séquence de *caractères*,

nous pouvons transposer facilement les méthodes de phylogénie.

Stemmatologie computationnelle

Les algorithmes basés sur la phylogénie peuvent être largement divisés en deux classes principales :

Stemmatologie computationnelle

Les algorithmes basés sur la phylogénie peuvent être largement divisés en deux classes principales :

- **Algorithmes basés sur la distance** : Calculent la distance entre chaque séquence et organisent les séquences les plus proches ensemble.

Stemmatologie computationnelle

Les algorithmes basés sur la phylogénie peuvent être largement divisés en deux classes principales :

- **Algorithmes basés sur la distance** : Calculent la distance entre chaque séquence et organisent les séquences les plus proches ensemble.
- **Algorithmes basés sur la probabilité** : Calculent pour chaque individu la probabilité d'être un descendant d'un autre.

Stemmatologie computationnelle

Les algorithmes basés sur la phylogénie peuvent être largement divisés en deux classes principales :

- **Algorithmes basés sur la distance** : Calculent la distance entre chaque séquence et organisent les séquences les plus proches ensemble.
- **Algorithmes basés sur la probabilité** : Calculent pour chaque individu la probabilité d'être un descendant d'un autre.
- **Algorithmes basés sur l'analyse sémantique**: ***approche nouvelle développée au cours du projet, cherchant à utiliser des outils de traitement automatique du langage naturel***.

Algorithmes basés sur la distance

Distance

Une **distance** est une fonction utilisée pour calculer la similarité entre des séquences de caractères.

Algorithmes basés sur la distance

Distance

Une **distance** est une fonction utilisée pour calculer la similarité entre des séquences de caractères.

Les distances possibles incluent :

- **Distances d'édition (Levenshtein)** : combien d'éditions faut-il pour muter une chaîne de caractères en une autre ? *Exemple* : d(*chien*, *chat*) = 1; d(*chien*, *rat*) = 2.

Algorithmes basés sur la distance

Distance

Une **distance** est une fonction utilisée pour calculer la similarité entre des séquences de caractères.

Les distances possibles incluent :

- **Distances d'édition (Levenshtein)** : combien d'éditions faut-il pour muter une chaîne de caractères en une autre ? *Exemple* : d(*chien*, *chat*) = 1; d(*chien*, *rat*) = 2.
- **Basées sur les tokens** : combien de tokens (= mots) sont en commun entre les phrases ? *Exemple* : d(*le chien est très mignon*, *le chat est très mignon*) = 1

Algorithmes basés sur la distance

Distance

Une **distance** est une fonction utilisée pour calculer la similarité entre des séquences de caractères.

Les distances possibles incluent :

- **Distances d'édition (Levenshtein)** : combien d'éditions faut-il pour muter une chaîne de caractères en une autre ? *Exemple* : d(*chien*, *chat*) = 1; d(*chien*, *rat*) = 2.
- **Basées sur les tokens** : combien de tokens (= mots) sont en commun entre les phrases ? *Exemple* : d(*le chien est très mignon*, *le chat est très mignon*) = 1
- **Distance sémantique** : à quelle distance sont les mots en termes de sens ? *Exemple* : d(*chien*, *chat*) > d(*chien*, *loup*)

Algorithmes basés sur la probabilité

Fonction de probabilité

Une fonction de probabilité est une fonction dans [0, 1] qui décrit à quel point un événement est probable (0, impossible, 1, certain).

Algorithmes basés sur la probabilité

Fonction de probabilité

Une fonction de probabilité est une fonction dans [0, 1] qui décrit à quel point un événement est probable (0, impossible, 1, certain).

Les algorithmes basés sur la probabilité nécessitent une fonction définissant à quel point il est *probable* qu'un manuscrit soit une copie d'un autre. Nous pouvons ensuite reconstruire chaque arbre et sélectionner **le plus probable**.

Exemple : $\mathbb{P}(\text{“Le chien est mignon”} \rightarrow \text{“Le chat est mignon”}) \geq (\text{“Le chien est mignon”} \rightarrow \text{“Le chien a mangé ma pizza”})$.

Algorithmes basés sur la probabilité

Plusieurs approches existent :

Algorithmes basés sur la probabilité

Plusieurs approches existent :

- **Moindres carrés** : Partir du principe que la prédiction de la longueur de l'arbre peut être faite en utilisant une régression par moindres carrés.

Algorithmes basés sur la probabilité

Plusieurs approches existent :

- **Moindres carrés** : Partir du principe que la prédiction de la longueur de l'arbre peut être faite en utilisant une régression par moindres carrés.
- **Maximum de vraisemblance** : Calculer chaque arbre et voir lequel a la plus grande vraisemblance (= est le plus probable étant donné le modèle de probabilité que nous avons choisi)

Sélection de l'algorithme approprié

Étant donné tous ces algorithmes possibles, **comment pouvons-nous sélectionner le bon** pour résoudre notre problème ?

Évaluation des algorithmes de stemmatologie

Évaluation

L'évaluation consiste à utiliser des métriques **objectives** (précision ...) pour évaluer les performances d'une méthode.

Valeur de référence (*ground truth*)

Une **valeur de référence (*ground truth*)** consiste en une réponse préalablement connue que l'algorithme est censé renvoyer.

Évaluation des algorithmes de stemmatologie

Évaluation

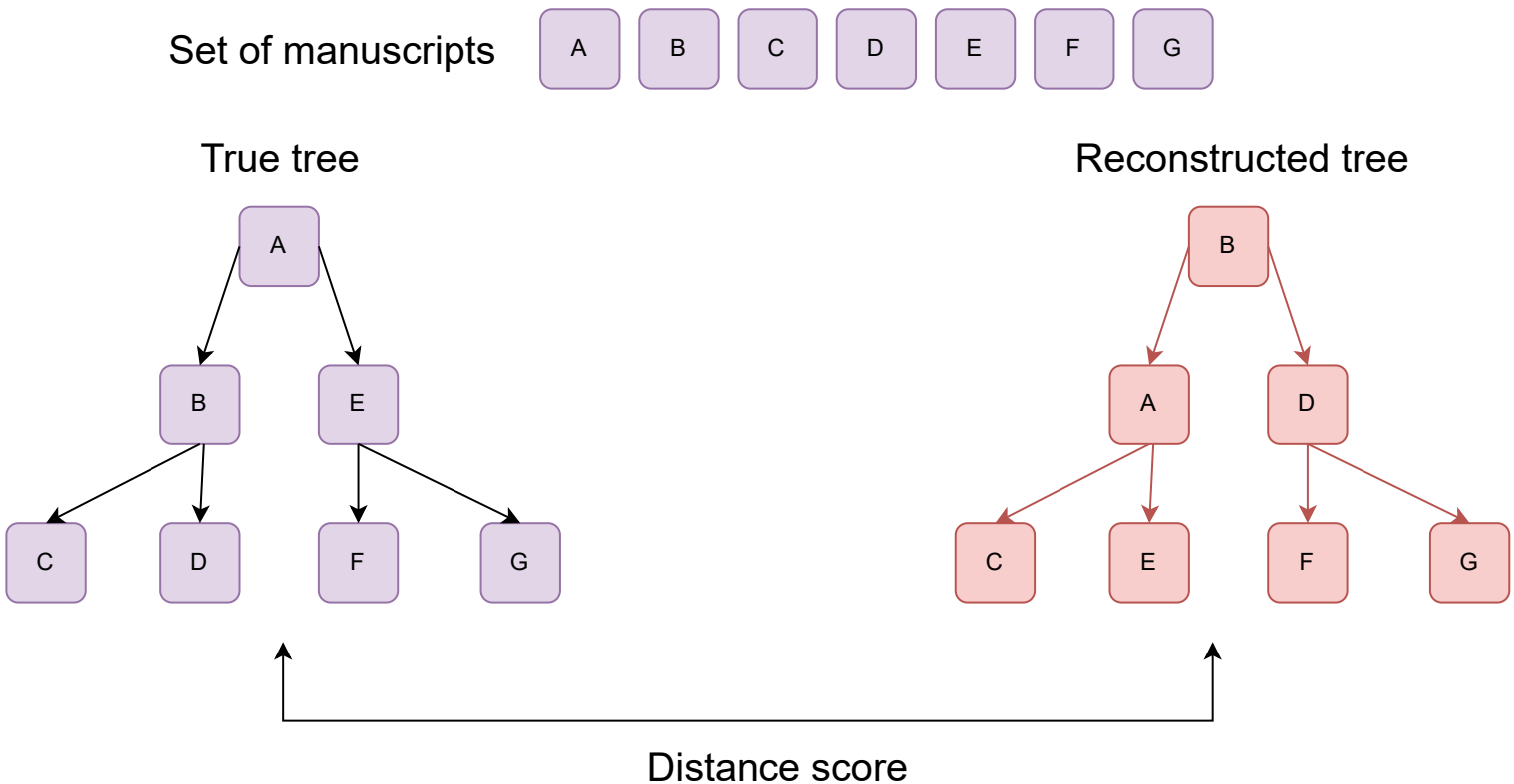
L'évaluation consiste à utiliser des métriques **objectives** (précision ...) pour évaluer les performances d'une méthode.

Valeur de référence (*ground truth*)

Une **valeur de référence (*ground truth*)** consiste en une réponse préalablement connue que l'algorithme est censé renvoyer.

L'une des principales caractéristiques et avantages de l'*apprentissage automatique* est que les performances des algorithmes peuvent être validées **de manière objective** en utilisant des métriques indépendantes.

Évaluation des algorithmes de stemmatologie



Évaluation des algorithmes de stemmatologie

Pour avoir une valeur de référence:

Évaluation des algorithmes de stemmatologie

Pour avoir une valeur de référence:

Deux approches possibles :

Évaluation des algorithmes de stemmatologie

Pour avoir une valeur de référence:

Deux approches possibles :

- Disposer de traditions écrites à la main avec une valeur de référence (*ground truth*) connue, qui peuvent être copiées artificiellement.

Évaluation des algorithmes de stemmatologie

Pour avoir une valeur de référence:

Deux approches possibles :

- Disposer de traditions écrites à la main avec une valeur de référence (*ground truth*) connue, qui peuvent être copiées artificiellement.
- Générer une tradition synthétique en utilisant des outils informatiques.

StemmaBench : génération de traditions artificielles

StemmaBench

StemmaBench est une bibliothèque pour la génération rapide de traditions manuscrites synthétiques.

Entrée : un texte, des paramètres de comportement scribal.
Sortie : une tradition manuscrite synthétique.

Example

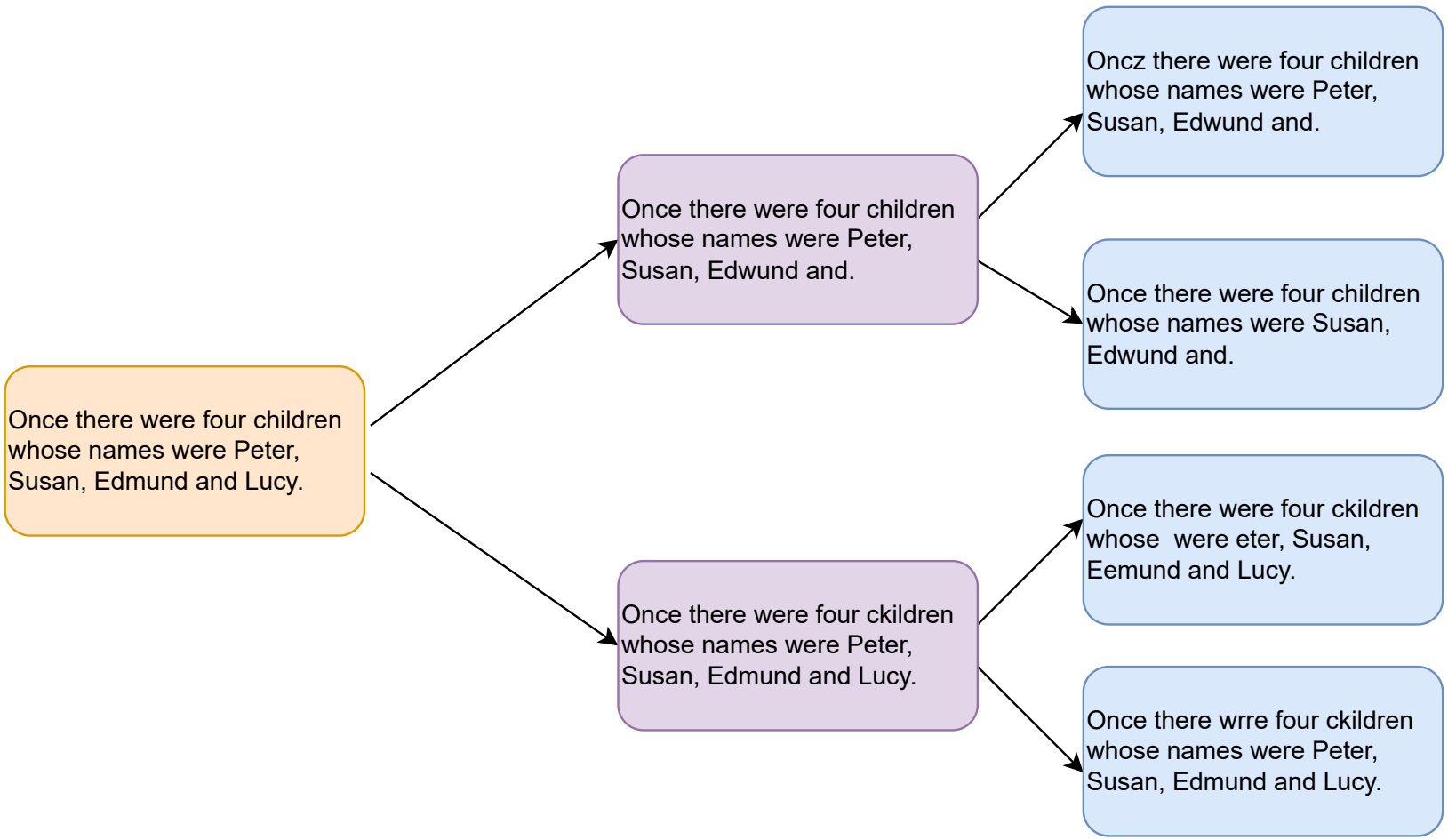
Input text: Extrait de *the Lion, The Witch and The Wardrobe*, by C.S. Lewis

> *Once there were four children whose names were Peter, Susan, Edmund and Lucy.*

Fichier de configuration:

Example

StemmaBench va permettre de générer la tradition suivante:



Prochaines étapes : amélioration de StemmaBench

Davide et moi travaillons sur:

- La prise en charge de l'hébreu biblique;

Prochaines étapes : amélioration de StemmaBench

Davide et moi travaillons sur:

- La prise en charge de l'hébreu biblique;
- La modélisation de la contamination;

Prochaines étapes : amélioration de StemmaBench

Davide et moi travaillons sur:

- La prise en charge de l'hébreu biblique;
- La modélisation de la contamination;
- L'expérimentation sur des textes bibliques.

Prochaines étapes : application à Qumrân

Les étapes de notre projet de recherche sont :

Prochaines étapes : application à Qumrân

Les étapes de notre projet de recherche sont :

1. L'analyse statistique des variantes trouvées à Qumran (**en cours, en collaboration avec Davide sur le projet BENTO :-)**);

Prochaines étapes : application à Qumrân

Les étapes de notre projet de recherche sont :

1. L'analyse statistique des variantes trouvées à Qumran (**en cours, en collaboration avec Davide sur le projet BENTO :-)**);
2. L'entrée de ces paramètres dans StemmaBench;

Prochaines étapes : application à Qumrân

Les étapes de notre projet de recherche sont :

1. L'analyse statistique des variantes trouvées à Qumran (**en cours, en collaboration avec Davide sur le projet BENTO :-)**);
2. L'entrée de ces paramètres dans StemmaBench;
3. L'évaluation des algorithmes de stemmatologie existants et le développement d'un nouveau sur ces traditions synthétiques;

Prochaines étapes : application à Qumrân

Les étapes de notre projet de recherche sont :

1. L'analyse statistique des variantes trouvées à Qumran (**en cours, en collaboration avec Davide sur le projet BENTO :-)**);
2. L'entrée de ces paramètres dans StemmaBench;
3. L'évaluation des algorithmes de stemmatologie existants et le développement d'un nouveau sur ces traditions synthétiques;
4. L'application des algorithmes les plus performants sur les données de Qumran ;

Prochaines étapes : application à Qumrân

Les étapes de notre projet de recherche sont :

1. L'analyse statistique des variantes trouvées à Qumran (**en cours, en collaboration avec Davide sur le projet BENTO :-)**);
2. L'entrée de ces paramètres dans StemmaBench;
3. L'évaluation des algorithmes de stemmatologie existants et le développement d'un nouveau sur ces traditions synthétiques;
4. L'application des algorithmes les plus performants sur les données de Qumran ;
5. La comparaison des stemmata générés automatiquement avec ceux générés manuellement.

Questions ?

Merci de votre attention :-)
Questions ?