# ALMA

## Automated Alignment of Ancient Texts using Linguistic and Semantic Analysis

Sophie Robert-Hayek

Variation linguistique et Linguistique Computationnelle

Sens Texte Informatique Histoire

09 Octobre 2025

# Introduction

# L'alignement et la philologie

Pour reconstruire les textes originaux et comprendre leur transmission :

- **Recensio** : catalogage des manuscrits disponibles
- **Collatio** : liste systématique des variantes intéressantes.

(Maas 1958)

# L'alignement et la philologie

Pour reconstruire les textes originaux et comprendre leur transmission :

- **Recensio** : catalogage des manuscrits disponibles
- **Collatio** : liste systématique des variantes intéressantes.

(Maas 1958)

| | | | | | | |
|---|---|---|---|---|---|---|
| ℵ | αρχη του ευαγγελιου | ιυ | χυ | | | |
| *A* | αρχη του ευαγγελιου | ιυ | χυ | υυ | του | θυ |
| *B* | αρχη του ευαγγελιου | ιυ | χυ | υιου | | θυ |
| *D* | αρχη του ευαγγελιου | ιηυ | χρυ | υιου | | θυ |
| $\theta$ | αρχη του ευαγγελιου | ιυ | χυ | | | |

# L'alignement et la philologie

Pour reconstruire les textes originaux et comprendre leur transmission :

- **Recensio** : catalogage des manuscrits disponibles
- **Collatio** : liste systématique des variantes intéressantes.

(Maas 1958)

| | | | | | | |
|---|---|---|---|---|---|---|
| ℵ | αρχη του ευαγγελιου | ιυ | χυ | | | |
| *A* | αρχη του ευαγγελιου | ιυ | χυ | υυ | του | θυ |
| *B* | αρχη του ευαγγελιου | ιυ | χυ | υιου | | θυ |
| *D* | αρχη του ευαγγελιου | ιηυ | χρυ | υιου | | θυ |
| *θ* | αρχη του ευαγγελιου | ιυ | χυ | | | |

¶ **1,1** ʻυιου του θεου (κυριου 1241) A K P Δ $f^{1.13}$ 33. 565. 579. 700. 892. 1241. 1424. 2542. *l* 844 𝔐 ¦ − ℵ* Θ 28. *l* 2211 $sa^{ms}$; Or (*et om.* Ιησου Χριστου Ir Epiph) ¦ *txt* ℵ¹ B D

# Dom Froger, le pionnier

- Chronophage et prône à l'erreur
- Lignes de transmission complexes entre plusieurs langues

# Dom Froger, le pionnier

Froger 1966 ; Froger 1968

- Chronophage et prône à l'erreur
- Lignes de transmission complexes entre plusieurs langues
- **Application parfaite pour les méthodes computationnelles**

# État de la recherche

- Les méthodes dominantes s'appuient sur l'alignement de séquences inspiré de la bio-informatique.
- Les récentes avancées en NLP dans l'alignement sémantique/syntaxique pour l'alignement de traduction sont sous-utilisées.

# État de la recherche

- Les méthodes dominantes s'appuient sur l'alignement de séquences inspiré de la bio-informatique.
- Les récentes avancées en NLP dans l'alignement sémantique/syntaxique pour l'alignement de traduction sont sous-utilisées.

$\rightarrow$ Pouvons-nous tirer parti des récentes avancées en matière de traduction automatique ?

# Formalisation mathématique

Si nous définissons :

- **Témoin de référence** : $(r_i)_{1 \leq i \leq n}$ (séquence de $n$ mots)
- **Témoin comparé** : $(c_j)_{1 \leq j \leq k}$ (séquence de $k$ mots)

Alignement représenté sous forme de matrice binaire $A \in \{0,1\}^{n \times k}$ :

$$a_{ij} = \begin{cases} 1 & \text{si le mot } r_i \text{ est aligné avec le mot } c_j \\ 0 & \text{sinon} \end{cases}$$

# Formalisation mathématique

Si nous définissons :

- **Témoin de référence** : $(r_i)_{1 \leq i \leq n}$ (séquence de $n$ mots)
- **Témoin comparé** : $(c_j)_{1 \leq j \leq k}$ (séquence de $k$ mots)

Alignement représenté sous forme de matrice binaire $A \in \{0,1\}^{n \times k}$ :

$$a_{ij} = \begin{cases} 1 & \text{si le mot } r_i \text{ est aligné avec le mot } c_j \\ 0 & \text{sinon} \end{cases}$$

- Alignements plusieurs-à-un autorisés
- Alignements un-à-plusieurs autorisés

# Exemple de matrice d'alignement

*Codex Palatinus/Codex Corbeiensis Secundus*

| | et | lux | lucet | in | tenebris | et | tenebrae | eum | non | comprae |
|---|---|---|---|---|---|---|---|---|---|---|
| et | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| lux | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| in | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| tenebris | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| lucet | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| et | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| tenebrae | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| eam | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| non | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| compraehenderunt | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

# État de l'art

# Algorithmes sequence-based

- Approche pionnière de Dom Froger.
- Comparaison mot à mot avec fenêtre glissante, inspirée de l'alignement des séquences protéiques (Robinson 1989)

✅ Indépendant de la langue
❌ Indépendant de la syntaxe et de la sémantique
❌ Peu adapté aux langues à ordre des mots libre (latin/grec)
❌ Ne peut pas gérer l'alignement interlinguistique

→ **Algorithme de Dekker dans CollateX.**

# Algorithmes syntax-based

Sultan, Bethard et Sumner 2014 :

- Utilise des outils NLP : lemmes, dépendances, entités nommées

# Algorithmes syntax-based

Sultan, Bethard et Sumner 2014 :

- Utilise des outils NLP : lemmes, dépendances, entités nommées
- Stratégie d'alignement :
  1. Aligner les séquences de mots similaires par lemme
  2. Utiliser les dépendances syntaxiques pour les mots restants
  3. Utiliser la base de données de paraphrases (PPDB) pour la similarité
  4. Aligner les mots vides avec des indices contextuels

✅ Comptes pour les lemmes et les formes fléchies

❌ Impossible d'aligner des mots sémantiquement équivalents en dehors du PPDB

# Algorithme basée sur le Deep Learning

- **Supervisé** : entraîner des modèles sur des données d'alignement (Stengel-Eskin et al. 2019 ; Nagata, Chousa et Nishino 2020 ; Zenkel, Wuebker et DeNero 2020)

# Algorithme basée sur le Deep Learning

- **Supervisé** : entraîner des modèles sur des données d'alignement (Stengel-Eskin et al. 2019 ; Nagata, Chousa et Nishino 2020 ; Zenkel, Wuebker et DeNero 2020)
- **Non supervisé** : apprentissage à partir de la cooccurrence des termes (Zenkel, Wuebker et DeNero 2020 ; Dou et Neubig 2021 ; Sabet et al. 2020) SimAlign Sabet et al. 2020 :
  - Méthode non supervisée maximisant la similarité dans l'espace d'intégration ;
  - Élagage via mesure d'entropie + critère de proximité ;
  - Ajustable sur des ensembles de données spécifiques.

# Algorithme basée sur le Deep Learning

- **Supervisé** : entraîner des modèles sur des données d'alignement (Stengel-Eskin et al. 2019 ; Nagata, Chousa et Nishino 2020 ; Zenkel, Wuebker et DeNero 2020)
- **Non supervisé** : apprentissage à partir de la cooccurrence des termes (Zenkel, Wuebker et DeNero 2020 ; Dou et Neubig 2021 ; Sabet et al. 2020) SimAlign Sabet et al. 2020 :
  - Méthode non supervisée maximisant la similarité dans l'espace d'intégration ;
  - Élagage via mesure d'entropie + critère de proximité ;
  - Ajustable sur des ensembles de données spécifiques.

✅ **Capacité sémantique** : Capable de capturer les variations sémantiques/orthographiques/grammaticales ;

❌ **Sur-correspondance** : tendance à la sur-correspondance due au calcul itératif de la distance ;

❌ **Syntaxe agnostique** : peut aligner différentes catégories grammaticales.

# Suggestion : ALMA

Nous proposons ALMA (**A**lignment and **L**earning for **M**anuscript **A**nalysis) pour pallier les limites actuelles en :

- Optimisant conjointement les alignements de lemmes, la cohérence grammaticale et la similarité sémantique
- **Exploitation des points forts** de chaque paradigme tout en atténuant les points faibles
- **Indépendance vis-à-vis de la langue**.

# Suggestion : ALMA

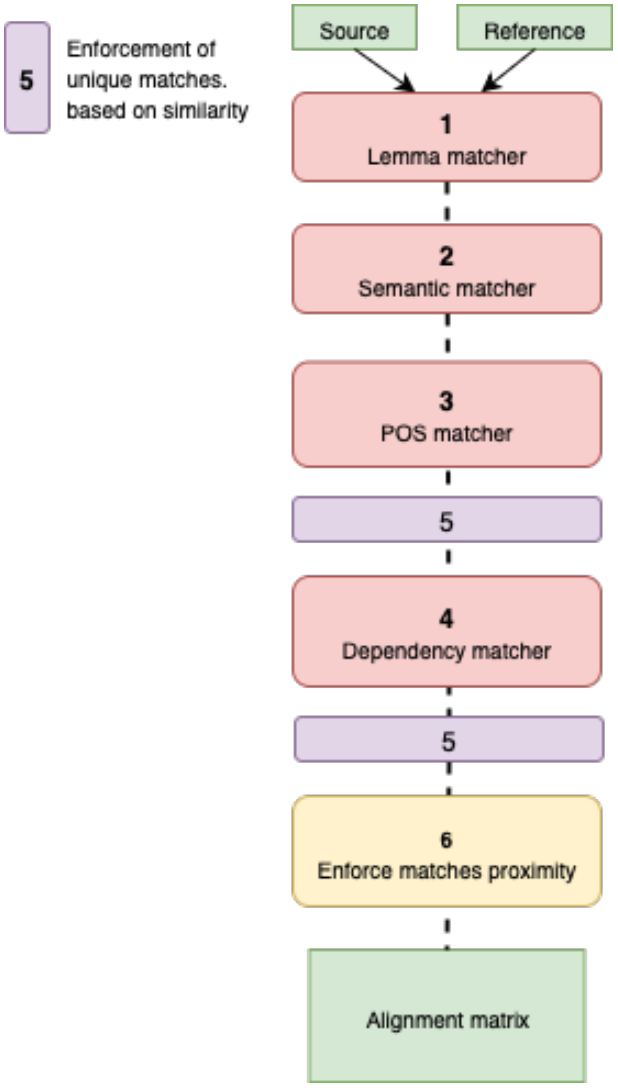
Nous proposons ALMA (**A**lignment and **L**earning for **M**anuscript **A**nalysis) pour pallier les limites actuelles en :

- Optimisant conjointement les alignements de lemmes, la cohérence grammaticale et la similarité sémantique
- **Exploitation des points forts** de chaque paradigme tout en atténuant les points faibles
- **Indépendance vis-à-vis de la langue**.

$\Rightarrow$ testés pour le mono-alignement de jeux de données grecs et latins.

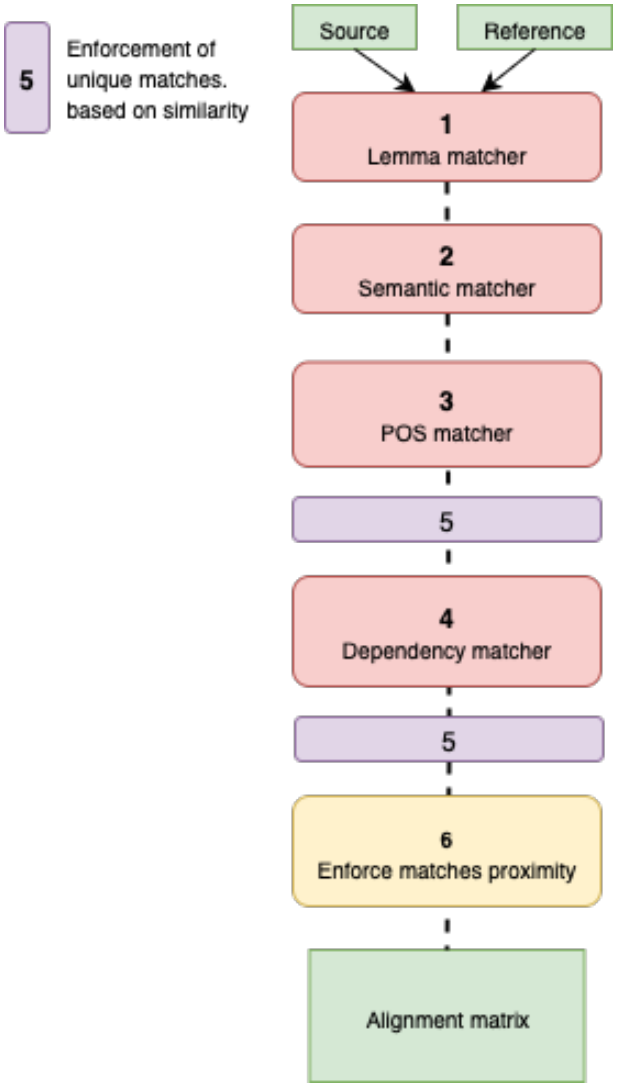
# ALMA

# Workflow




- **Matching** :
  - Identité de lemmes ;
  - Proximité sémantique ;
  - Identité de Partie du Discours ;
  - Identité de dépendance grammaticale ;

# Workflow




- **Matching** :
  - Identité de lemmes ;
  - Proximité sémantique ;
  - Identité de Partie du Discours ;
  - Identité de dépendance grammaticale ;
- **Élagage** :
  - Unicité de la correspondance basée sur la proximité sémantique ;
  - Distance entre les correspondances dans la référence/comparé.

# Détecter la synonymie

**Similitude cosinus** : $S_{ij} = \frac{\mathbf{r}_i \cdot \mathbf{c}_j}{\|\mathbf{r}_i\| \|\mathbf{c}_j\|}$
**Normalisation du score Z** :
$z_{ij} = \frac{S_{ij} - \mu_i}{\sigma_i}$
**Sélection des correspondances** :
$l^* = \arg\max_l z_{il}$ si $z_{il^*} > \tau$

On :

- Calcule la matrice de similarité ;
- Normalise par ligne avec le *z-score* ;
- Sélectionne les correspondances où la similarité se démarque du contexte de la phrase (c'est-à-dire $> \tau$)

# Filtre d'unicité

**Sélection par ligne** :
$j^* = \arg\max_j S_{ij}$
**Sélection par colonne** :
$i^* = \arg\max_i S_{ij}$

- Appliqué après le POS et la correspondance de dépendance
- En cas de correspondances multiples, sélectionne la similarité sémantique la plus élevée
- **Garantit que les correspondances grammaticales respectent le contexte sémantique**
- Empêche le suralignement des mots fonctionnels

# Distortion filter

En définissant la différence de position normalisée comme suit :

$$\delta_{ij} = \frac{i}{n} - \frac{j}{k}$$

$$|\delta_{ij}| > \tau \Rightarrow A_{ij} = 0$$

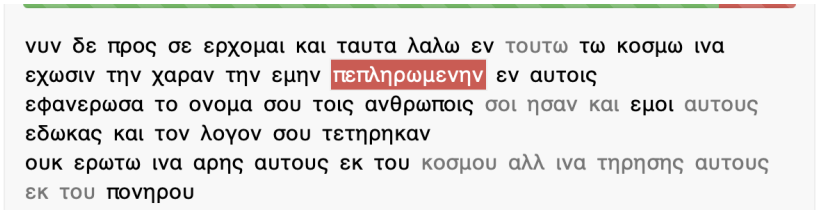
- Garantit que les mots alignés apparaissent dans des positions relatives similaires.
- Le seuil $\tau \in [0, 1]$ supprime les alignements trop éloignés.
- Offre une alternative plus simple au modèle de distorsion IBM 2.

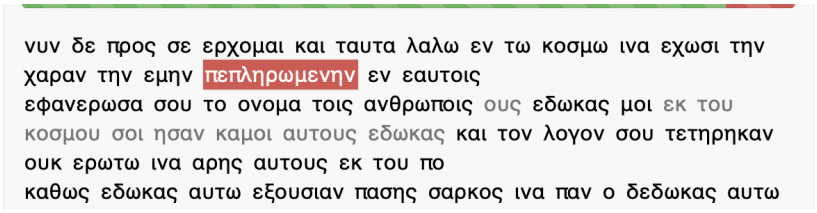
# Évaluation et résultats

# Ensembles de données d'évaluation

L'évaluation a été réalisée sur deux **ensembles de données d'alignement monolingues** en latin et en grec, avec un échantillonnage aléatoire de 15 versets sur 19 chapitres de l'Évangile selon Jean.

- *Vetus Latina* : Codex Corbeiensis Secundus (européen) vs Codex Palatinus (africain) ;
- *Nouveau Testament grec* : Codex Bezae (occidental) et Vaticanus (alexandrin).



# Ground truth

| | Latin | Grec |
|---|---|---|
| **Total des tokens** | 4880 | 5078 |
| **Tokens avec correspondances** | 87,9 % | 89,9 % |
| **Tokens sans correspondances** | 12,1 % | 10,1 % |
| **Correspondances non identiques** | 25,1 % | 10,6 % |

- L'ensemble de données en latin offre une plus grande **liberté** textuelle.
- L'ensemble de données en grec ne comporte pas de signes diacritiques, **invisibles pour le lemmatiseur/étiqueteur POS**.

# Evaluation Metrics

Och et Ney 2000, **Alignment Error Rate** :

$$\text{AER}(A,S) = 1 - \frac{2 \times |A \cap S|}{|A| + |S|}((1 - \text{AER}) \times 100)$$

- **Précision** : $\frac{|A \cap S|}{|A|}$ - évite les liens incorrects
- **Rappel** : $\frac{|A \cap S|}{|S|}$ - trouve des alignements valides

# Métriques d'évaluation

$$\text{AER}_{\text{diff}} = \frac{2\,|A_{\text{diff}} \cap T_{\text{diff}}|}{|A_{\text{diff}}| + |T_{\text{diff}}|} \times 100$$

- $T_{\text{diff}}$ : paires de mots différentes de référence où $r_i \neq c_j$
- $A_{\text{diff}}$ : alignements prédits pour $r_i$

$\rightarrow$ Nuance les scores élevés liés aux nombreux mots identiques.

# Modèles neuronaux

- **Taggers** :
  - **Latin** : pipeline LatinCy spaCy
  - **Grec** : modèle Stanza personnalisé entraîné sur PROIEL sans accents ;
- **Embeddings** : mERT/XLM-RoBERTa :
  - Modèle « brut » ;
  - Modèle fine-tuned (MLM) sur collection de manuscrits en latin et en grec.

# Global alignment scores : AER

**Latin**

| | Model | AER |
|---|---|---|
| **ALMA** | RoBERTa-tuned | 93.80 |
| | mBERT-tuned | 93.37 |
| | mBERT | 92.88 |
| | RoBERTa | 92.48 |
| **SimAlign** | RoBERTa-tuned | 92.33 |
| | RoBERTa | 91.24 |
| | mBERT | 89.57 |
| | mBERT-tuned | 88.81 |
| **POS** | | 76.29 |
| **Collatex** | LevDist = False | 67.93 |
| | LevDist = True | 67.93 |

**Greek**

| | Model | AER |
|---|---|---|
| **ALMA** | mBERT-tuned | 96.90 |
| | mBERT | 96.57 |
| | XLM-R-tuned | 96.29 |
| | XLM-R | 95.33 |
| **SimAlign** | XLM-R-tuned | 94.74 |
| | XLM-R | 93.79 |
| | mBERT | 93.71 |
| | mBERT-tuned | 91.78 |
| **POS** | | 86.95 |
| **Collatex** | LevDist = False | 84.46 |
| | LevDist = True | 84.46 |

# Alignment Results : Latin & Greek

- **ALMA surpasse tous ses concurrents** dans les deux langues :
  - **Latin** : +1,47 point AER
  - **Grec** : +2,16 points AER
- **Amélioration considérable par rapport aux méthodes traditionnelles** :
  - Avantage de plus de +20 points AER par rapport à POS et Collatex dans les deux langues
- **Importance du fine-tuning** :
  - RoBERTa-tuned offre les meilleures performances en latin
  - mBERT-tuned meilleur

# Global alignment scores : AERDiff

**Latin Texts**

| | Model | AER Diff |
|---|---|---|
| **ALMA** | RoBERTa-tuned | 87.39 |
| | BERT-tuned | 85.53 |
| | mBERT | 84.83 |
| | XLM-R | 81.96 |
| **SimAlign** | RoBERTa-tuned | 84.75 |
| | XLM-R | 82.44 |
| | mBERT | 73.39 |
| | BERT-tuned | 73.60 |
| **POS** | | 32.07 |
| **Collatex** | | 0.00 |

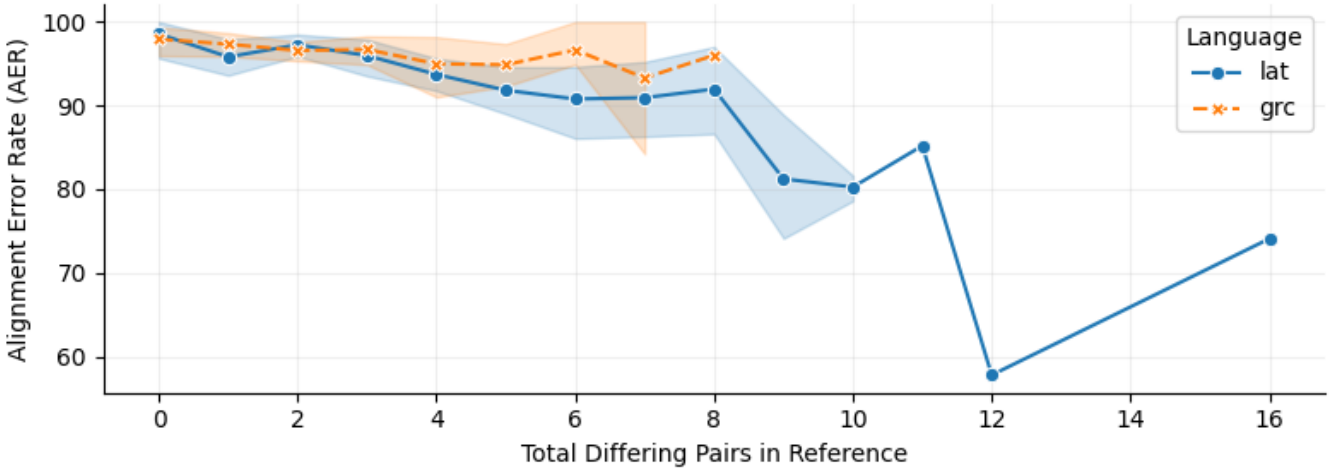
**Greek Texts**

| | Model | AER Diff |
|---|---|---|
| **ALMA** | BERT-tuned | 88.66 |
| | mBERT | 87.86 |
| | RoBERTa-tuned | 86.08 |
| | XLM-R | 80.39 |
| **SimAlign** | RoBERTa-tuned | 81.91 |
| | XLM-R | 76.37 |
| | mBERT | 77.43 |
| | BERT-tuned | 70.34 |
| **POS** | | 10.52 |
| **Collatex** | | 0.00 |

# Global alignment scores : AERDiff

- **Ne permet pas que le matching exact** :
  - **Latin** : 87,39 AERDiff
  - **Grec** : 88,66 AERDiff
- **Avantage substantiel par rapport aux méthodes concurrentes** :
  - **Latin** : +2,64 points par rapport au meilleur modèle SimAlign
  - **Grec** : +6,75 points par rapport au meilleur modèle SimAlign
  - Amélioration considérable par rapport aux approches basées sur des règles (POS, Collatex)
- **Intérêt des approches neuronales** :
  - Capture les relations sémantiques malgré les différences ;
  - Surpasse les méthodes reposant uniquement sur la similarité des chaînes ou les balises POS.

# Impact of token difference on ALMA performance




Malgré une bonne résiliance aux tokens différences, baisse de l'AER sur les versets avec beaucoup de pairs différentes.

# Cas 1

*Jn 5 :35 (Vetus Latina)* :
**Source :** *ille fuit lucerna ardens et lucens uos autem uoluistis exultari ad oram in lumine eius* [he was a lamp burning and shining, but you wished to be exulted for an hour in his light]
**Target :** *ille erat lucerna ardens et lucet lucens uos autem uoluistis ad horam exultare in luce eius* [he was a lamp burning and he shines shining, but you wished for an hour to exult in his light]

# Cas 1

*Jn 5 :35 (Vetus Latina) :*
**Source :** *ille fuit lucerna ardens et lucens uos autem uoluistis exultari ad oram in lumine eius* [he was a lamp burning and shining, but you wished to be exulted for an hour in his light]
**Target :** *ille erat lucerna ardens et lucet lucens uos autem uoluistis ad horam exultare in luce eius* [he was a lamp burning and he shines shining, but you wished for an hour to exult in his light]

| Algorithm | Matches | Unmatches |
|---|---|---|
| **ALMA** | ille ↔ ille, fuit ↔ erat, lucerna ↔ lucerna, ardens ↔ ardens, et ↔ et, lucens ↔ lucens, vos ↔ vos, autem ↔ autem, uoluistis ↔ uoluistis, ad ↔ ad, oram ↔ horam, exultari ↔ exultare, in ↔ in, lumine ↔ luce, eius ↔ eius | - ↔ lucet |
| **SimAlign** | ille ↔ ille, fuit ↔ erat, lucerna ↔ lucerna, ardens ↔ ardens, et ↔ et, **lucens ↔ lucet**, lucens ↔ lucens, vos ↔ vos, autem ↔ autem, uoluistis ↔ uoluistis, exultari ↔ exultare, ad ↔ ad, oram ↔ horam, in ↔ in, lumine ↔ lucet, lumine ↔ luce, eius ↔ eius | (none) |

# Case 2

*Jn 1 :13 (NT Greek)* :
**Source** : οὐκ ἐξ αἵματος οὐδὲ ἐκ θελήματος σάρκος οὐδὲ θελήματος ἀνδρός ἀλλὰ ἐκ θεοῦ ἐγεννήθησαν [Not from blood, nor from the will of the flesh, nor from the will of man, but from God they were born]
**Target** : οἱ οὐκ ἐξ αἵματος οὐδὲ ἐκ θελήματος σάρκος ἀλλὰ ἐκ θεοῦ ἐγενήθησαν [Those that were not born from blood, nor from the will of the flesh, but from God]

| Algorithm | Matches | Unmatches |
|---|---|---|
| ALMA | ουκ ↔ ουκ, εξ ↔ εξ, αἵματος ↔ αἵματος, ουδε ↔ ουδε, εκ ↔ εκ, θεληματος ↔ θεληματος, σαρκος ↔ σαρκος, αλλα ↔ αλλα, εκ ↔ εκ, θεου ↔ θεου, εγεννηθησαν ↔ εγενηθησαν | θεληματος ↔ -, ανδρος ↔ -, ουδε ↔ -, - ↔ οι |
| SimAlign | ουκ ↔ ουκ, εξ ↔ εξ, αἵματος ↔ αἵματος, ουδε ↔ ουδε, εκ ↔ εκ, θεληματος ↔ θεληματος, σαρκος ↔ σαρκος, **ουδε ↔ ουδε**, **θεληματος ↔ αἵματος**, **θεληματος ↔ θεληματος**, αλλα ↔ αλλα, εκ ↔ εκ, θεου ↔ θεου, εγεννηθησαν ↔ εγενηθησαν | ανδρος ↔ -, - ↔ οι |

# Case 3

**Source :** *dum sum in seculo lux sum seculi* [while I am in the world, I am the light of the world]
**Target :** *cum in hoc mundo sum lux sum huius mundi* [when in this world I am the light of this world]

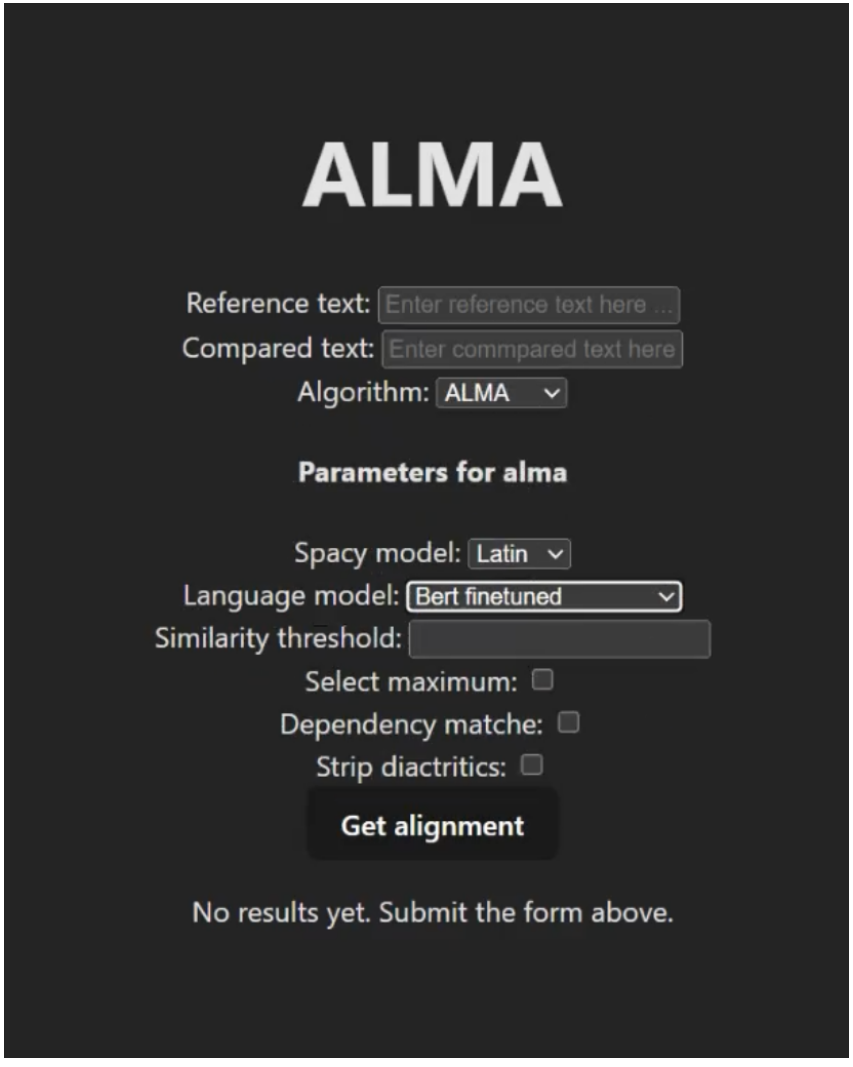
| Algorithm | Matches | Unmatches |
|---|---|---|
| **ALMA** | dum ↔ cum, sum ↔ sum, in ↔ in, seculo ↔ mundo, lux ↔ lux, sum ↔ sum, seculi ↔ mundi | - ↔ hoc<br>- ↔ huius |
| **SimAlign** | dum ↔ cum, sum ↔ sum, in ↔ in, **seculo ↔ hoc**, seculo ↔ mundo, lux ↔ lux, sum ↔ sum, **seculi ↔ huius**, seculi ↔ mundi | (none) |

# Conclusions and perspectives

# Python interface

```python
from alma import Collation

collator = Collation(collation_model="collatex")

collator = Collation(collation_model="pos",
                     model_language="latin")

collator = Collation(collation_model="simalign",
                     matching_methods="a", simalign_model="bert-base-multilingual-cased")

collator = Collation(collation_model="alma",
                     spacy_model= "la_core_web_lg",
                     similarity_threshold=1,
                     embedding_model="google-bert/bert-base-multilingual-cased" )

# Perform the collation
collation_matches = collator.collate(
    "A",
    "extersit pedes eius capillis suis cuius frater lazarus infirmabatur",
    "B",
    "exterserat capillis suis pedes eius cuius frater lazarus infirmabatur")

# Print the results of the collation
print(collation_matches)
```

# Web interface



# Conclusions

## Conclusion générale

- ALMA démontre des performances très intéressantes pour toutes les langues classiques ;
- Surpasse les approches existantes ;
- **Ouvre la voie à l'intégration des réseaux neuronaux pour la tâche d'alignement en philologie.**

# Perspectives : améliorations

- Étude complète sur l'ablation/les hyperparamètres ;
- Évaluation de la collation multilingue (grec/latin) ;
- Déploiement public :-)

# Perspectives : Applications

- **Philologie computationnelle :**
  - Évaluation de l'importance des variantes ;
  - Génération automatique de l'*appareil critique* ;

# Perspectives : Applications

- **Philologie computationnelle :**
  - Évaluation de l'importance des variantes ;
  - Génération automatique de l'*appareil critique* ;
- **Transmission textuelle :**
  - Comprendre la transmission du *Vetus Latina* des 2 Maccabées.
  - Ajouter le syriaque et le grec ?

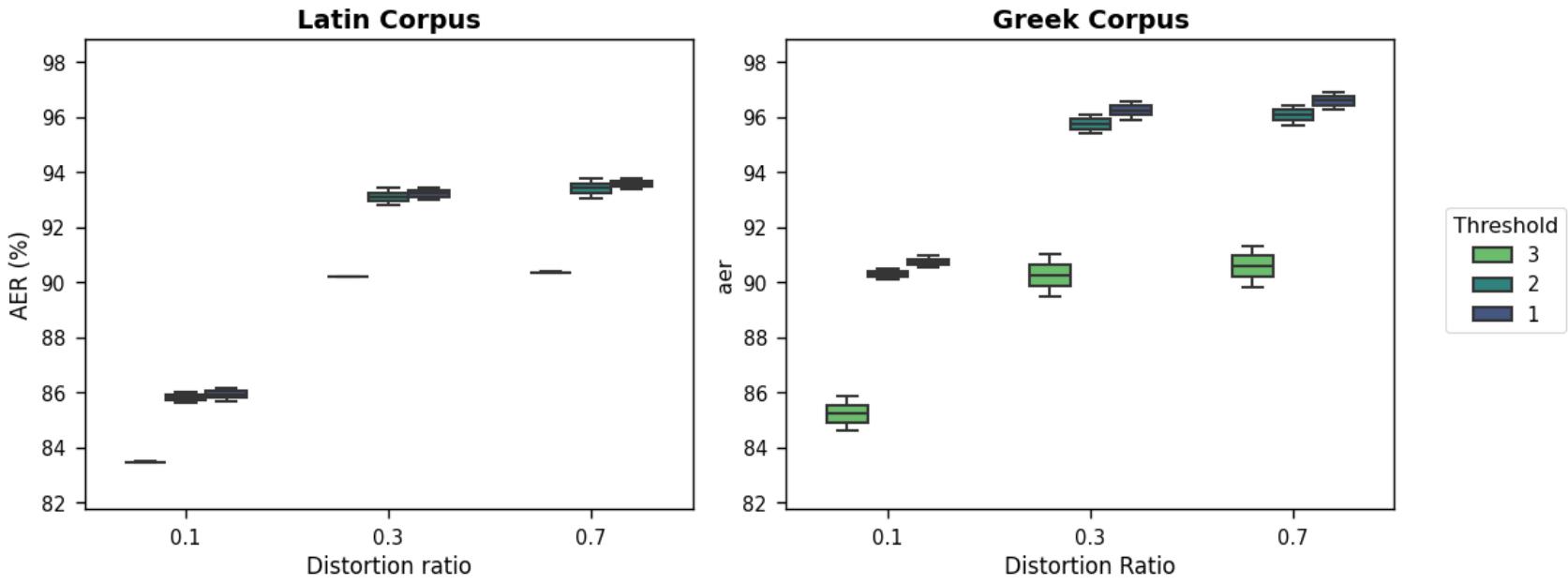
# Questions ?

Questions ?

# Ablation study - Preliminary data

| Type | Variant | AER (L) | AER_diff (L) | AER (G) | AER_diff (G) |
|---|---|---|---|---|---|
| **Full pipeline** | Baseline | 93.80 | 86.77 | 96.90 | 87.71 |
| **Component Removal** | -Lemma | 93.93 | 86.88 | 97.00 | 88.95 |
| | -Semantic | 91.19 | 80.08 | 88.30 | 62.33 |
| | - (POS + Dep) | 91.60 | 82.43 | 94.94 | 78.93 |
| **Order Change** | Semantic→Lemma | 94.18 | 87.54 | 96.84 | 88.20 |
| **Post-processing** | -Uniqueness | 82.39 | 68.76 | 86.04 | 63.03 |
| | -Proximity | 93.94 | 87.41 | 96.83 | 87.62 |

# Impact of hyperparameters




- **Threshold** :
  - Too strict threshold prevents matching.
  - Optimum seems around 1.
- **Distortion** :
  - Distortion ratio seems to do more harm than good.