Variation graphique dans les documents d'Ancien Régime : Nouvelles approches scriptométriques

Simon Gabay, Philippe Gambette, Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Benoît Sagot

Variation graphique dans les documents d'Ancien Régime

Nouvelles approches scriptométrique

Simon Gabay[1] Philippe Gambette[2] Rachel Bawden[3] Eleni Kogkitsidou[2]
Jonathan Poinhos[2] Benoît Sagot[3]

[1]Université de Genève (Suisse) – prenom.nom@unige.ch
[2]LIGM, Université Gustave Eiffel (France) – prenom.nom@u-pem.fr
[3]INRIA (France) – prenom.nom@inria.fr

Au commencement était la marquise

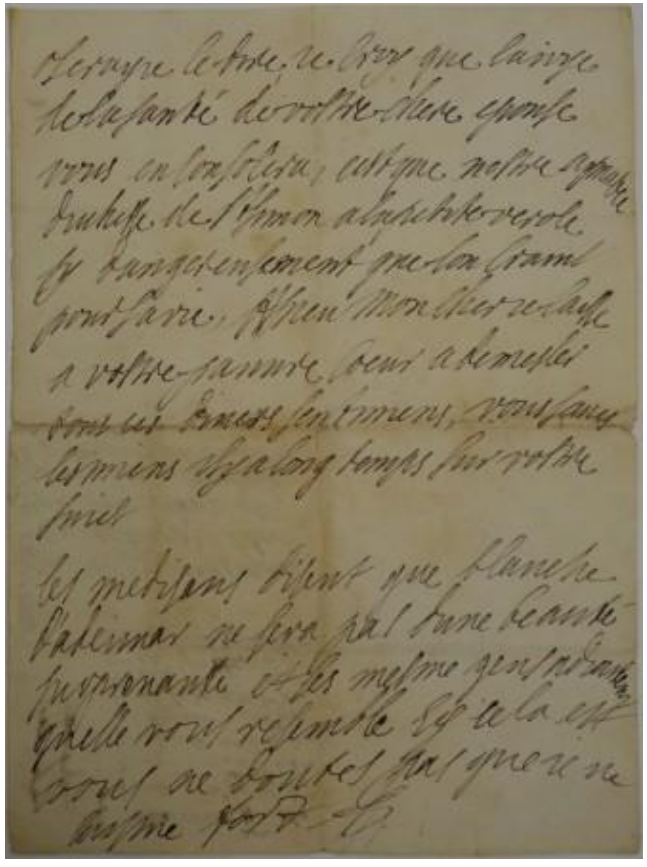


Figure 1 – Lettre de Sévigné/Grignan, MS Lowell Autograph File 282, f°3r.

- Tout a commencé avec la correspondance de la marquise
- Problème pour différencier les parties écrites par sa mère et sa fille
- Usage de l'accent, distinction <i>/<j> et <u>/<v>, usage de l'apostrophe, usage de la ponctuation…
- Très peu de travaux sur le français « classique »

[Simon Gabay, «La naissance de Marie-Blanche de Grignan. Notes sur la mise en page de la polyphonie sévignéenne», *Acta Litt&Arts* [En ligne], *Acta Litt&Arts, Les discours rapportés en contexte épistolaire (XVI^e^-XVIII^e^ s.)*]

Plan de la présentation

Des *scriptae* en français classique

De l'utilité de la scriptométrie

Approche par règles

Approche par réseau de neurones

Conclusion

Des *scriptae* en français classique

Des *scriptae* françaises ?

Deux grands courants orthographiques (Pellat 1991)

- Courant de l'« orthographe ancienne », apparu vers le XIII^{e} s., qui cherche à maintenir le lien avec le latin
- Courant de l'« orthographe nouvelle », qui cherche un système propre au français (et plus ancien que l'orthographe ancienne

Réalisation de cette opposition en deux *scriptae* au XVII^{e} s. :

- Par analogie avec les *scriptae* romanes, dérivées spontanément du latin et organisées diatopiquement (wallon, picard, anglo-normand
- Nous proposons de parler de *scriptae* françaises, dérivées de la norme médiévale et organisées diastratiquement sur le modèle de la Querelle des Anciens et des Modernes

[J.-Ch. Pellat, « Norme et variation orthographique au XVII^{e} s. », *Rencontres linguistiques en pays rhénan 5/6*, 1994, pp. 245-260.]

Exemple de Bossuet : théorie

Le choix du système graphique est théorisé par nombre d'auteurs, et influencé par leur position dans la Querelle.

> *[La compaignie] ne peut souffrir une fausse regle qu'on a uoulu introduire, d'écrire comme on prononce, parce qu'en uoulant instruire les étrangers et leur faciliter la prononciation de nostre langue, on la fait meconnoistre aux François mesmes. Si on ecrivoit tans, chan, cham, emais ou émés, connoissans, anterreman, faisaict, qui reconnoistroit ces mots ? [...] Il y a aussi une autre ortographe, qui s'attache scrupuleusement a toutes les lettres tirées des langues dont la nostre a pris ses mots, et qui ueut ecrire nuict, ecripture, etc. Cella blesse les yeux d'une autre sorte en leur remettant en ueüe des lettres dont ils sont desaccoustumez et que l'oreille n'a iamais connus.[Bossuet, Cahiers de Mezeray]*

[L. Biederman-Pasques, *Les Grands Courants orthographiques au XVII^e s. et la formation de l'orthographe moderne*, Tübingen : Max Niemeyer Verlag, 1992]

scripta des Anciens vs *scripta* des Modernes

-	Anciens	Modernes
Lettres ramistes	Position (*vniuers*)	Dissimilation (*univers*)
Ancien hiatus	*eu* (*veu*)	*u* (*vu*) ou *û* (*vû*)
Lettre calligraphique	maintenue (*ay*)	supprimée (*ai*)
Pluriel nominal	*-s*, *-z*, *-x*	*-s*
Consonne muette (diac., étym, hist.)	maintenue (*doubte*)	supprimée (*doute*)
Voyelles longues	diacritique (*teste*) dédoublement (*aage*)	circonflexe (*tête*)
...	...	...

Table 1 – Exemples d'opposition entre graphie ancienne et graphie moderne

- le système alphabétique, avec les couples <i>/<j> et <u>/<v> étant conçu comme des variantes graphiques ou comme renvoyant à des phonèmes différents
- les logogrammes lexicaux, avec l'utilisation ou la suppression de lettres en surcharge, qu'elles soient historiques (*huile*) ou étymologiques (*doubter*)
- les signes auxiliaires, avec l'utilisation de lettres diacritiques (*hospital*) plutôt que d'accents ou de trémas.

Exemple de Bossuet : pratique

Opposition relativement nette entre :

- Bossuet moderne, avant la fin de ses études (doctorat en 1652)
- Bossuet ancien, après avoir rejoint les ordres en Lorraine (1654)

Manuscrit	tant	être	cette	même	paraître	avec	a-t-il
BNF Fr. 12822, f°370 *Brièveté de la vie*, 1648	tans	estre	ceste	mesme	parêtre	auecque	a t'il
BNF Fr. 12823, f°130 *Fête du Rosaire*, oct. 1651	tans	être	cête	même	parêtra	auec	a t'il
BNF Fr. 12824, f°119 *Sur la Prov.*, mai 1656	temps	estre	ceste	mesme	paroist	auec/ auecque	a til
BNF Fr. 12822, f°61 *Sur les démons*, fév. 1660	temps	estre	cette	mesme	paroistre	-	atil

Table 2 – Exemples de l'évolution du système graphique de Bossuet

→ Grande richesse d'une analyse graphématique du point de vue linguistique, mais aussi littéraire et historique

[Bossuet, *Œuvres oratoires*, ed. Joseph Lebarq, Lille et Paris : Desclée, De Brouwer et Cie, 1890]

Plan de la présentation

Des *scriptae* en français classique

De l'utilité de la scriptométrie

Approche par règles

Approche par réseau de neurones

Conclusion

De l'utilité de la scriptométrie

Pourquoi une approche computationnelle ?

Quelle est l'utilité d'une approche computationnelle ?

- L'apparition et l'accélération des besoins numériques nécessite un outillage de qualité
  - La qualité des outils (lemmatisation, annotations diverses...) dépend de notre connaissance de la langue
  - Mais, à rebours, les outils peuvent aussi servir à améliorer notre connaissance de la langue
- Il existe un besoin de produire des textes normalisés (lecture, requête dans un corpus...)
  - La normalisation repose sur des règles (computationnelles) de transformation
  - Il devrait être possible d'étudier l'application de ces règles pour analyser la langue des textes

Deux corpus

Ces approches computationelles « *data-driven* » nécessitent cependant la création d'au moins deux corpus :

- Un corpus « gold » d'entraînement et d'évaluation des modèles de normalisation, proposant le même corpus dans une version normalisée ou non. → Parallel17
- Un corpus de recherche, sur lequel appliquer les outils développés avec le corpus « gold ». → d'AlemBERT

Parallel17

L'approche la plus logique est celle d'une apprentissage supervisée : l'ordinateur doit déduire des règles à partir d'exemples. À cette fin, nous avons conçu le corpus PARALLEL17 (plus de 650 000 tokens)

- On trouve une cinquantaine de textes du XVII^{e} s., parfois complets (*Andromaque* de Racine), parfois non (*Astrée* d'Urfé).
- le corpus se veut représentatif, et contient du vers comme de la prose, différents genres (surtout littéraires), est raisonnablement distribué diachroniquement par décennie...
- Chaque texte est découpé segments constitués de (sous-)phrases (délimiteurs retenus : <;>, < :>, <?>, <!>, <.>)
- Chaque segment offre une transcription diplomatique et sa version normalisée, alignée sur le français contemporain à quelques exceptions près (*i.e.* licence poétique comme l'élision pour raison métrique).

The bigger, the better

En plus du corpus d'entraînement, il nous faut un corpus de recherche

- Nous sommes confrontés à un problème important : celui de la normalisation quasi intégrale des éditions de texes classiques
- Il existe des corpus plus ou moins accessibles (et plutôt moins que plus), du type du RCFC, pas forcément utilisables pour des analyses graphématiques (type corpus Vachon, amputé pour le XVII^{e} s.)
- L'apparition de nouvelles technologies de numérisation et d'analyse amène l'utilisation de corpus de très grande taille

d'AlemBERT

Nous nous sommes donc lancés dans la constitution d'une super corpus du français moderne

- Couvre la période XVI^{e} s.-XVIII^{e} s.
- Avec des données riches et harmonisées (auteur, date, genre, normalisation…)
- Nous avoisinons les 300 Mo de texte, mais il devrait continuer de grandir
- Problème de distribution des données et de publication sous forme de ressource

Ce corpus devrait permettre l'entraînement de modèles de langue dont l'objectif est de venir en soutient au tâches de TAL, notamment la normalisation

Plan de la présentation

Des *scriptae* en français classique

De l'utilité de la scriptométrie

Approche par règles

Approche par réseau de neurones

Conclusion

Approche par règles

Détection automatique de *scripte*

Approche fondée sur l'alignement de corpus parallèles

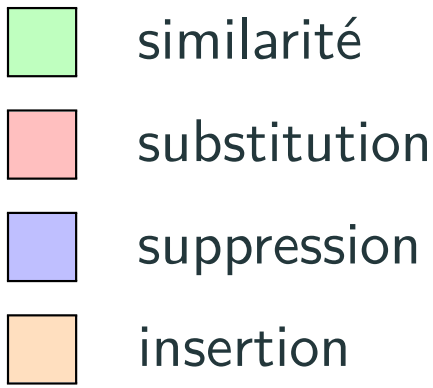
- Analyse scriptologique intégrée à la tâche de normalisation linguistique
- Utilisation du corpus parallèle Parallel17 comme base d'apprentissage
- Développement de la méthode ABA (*Alignement-Based Approach*) :
  - alignement au niveau des mots pour chaque (sous-)phrase de Parallel17
  - alignement au niveau des caractères pour chaque mot
  - pour chaque caractère modifié entre la version originale et normalisée, détection de la règle appliquée
  - Utilisation de ces règles pour normaliser un texte
  - La règle utilisée est sauvegardée pour une analyse linguistique

Alignement avec l'algorithme de Needleman-Wunsch

L'algorithme de Needleman-Wunsch est un algorithme qui effectue un alignement global maximal de deux chaînes de caractères. On utilise pour cela une matrice de similarité

	A	p	o	ſ	t	r	e
A							
p							
ô							
t							
r							
e							

Table 3 – Matrice de similarité pour la version originale et normalisée d'*Apoſtre*.



Les règles dans la méthode ABA

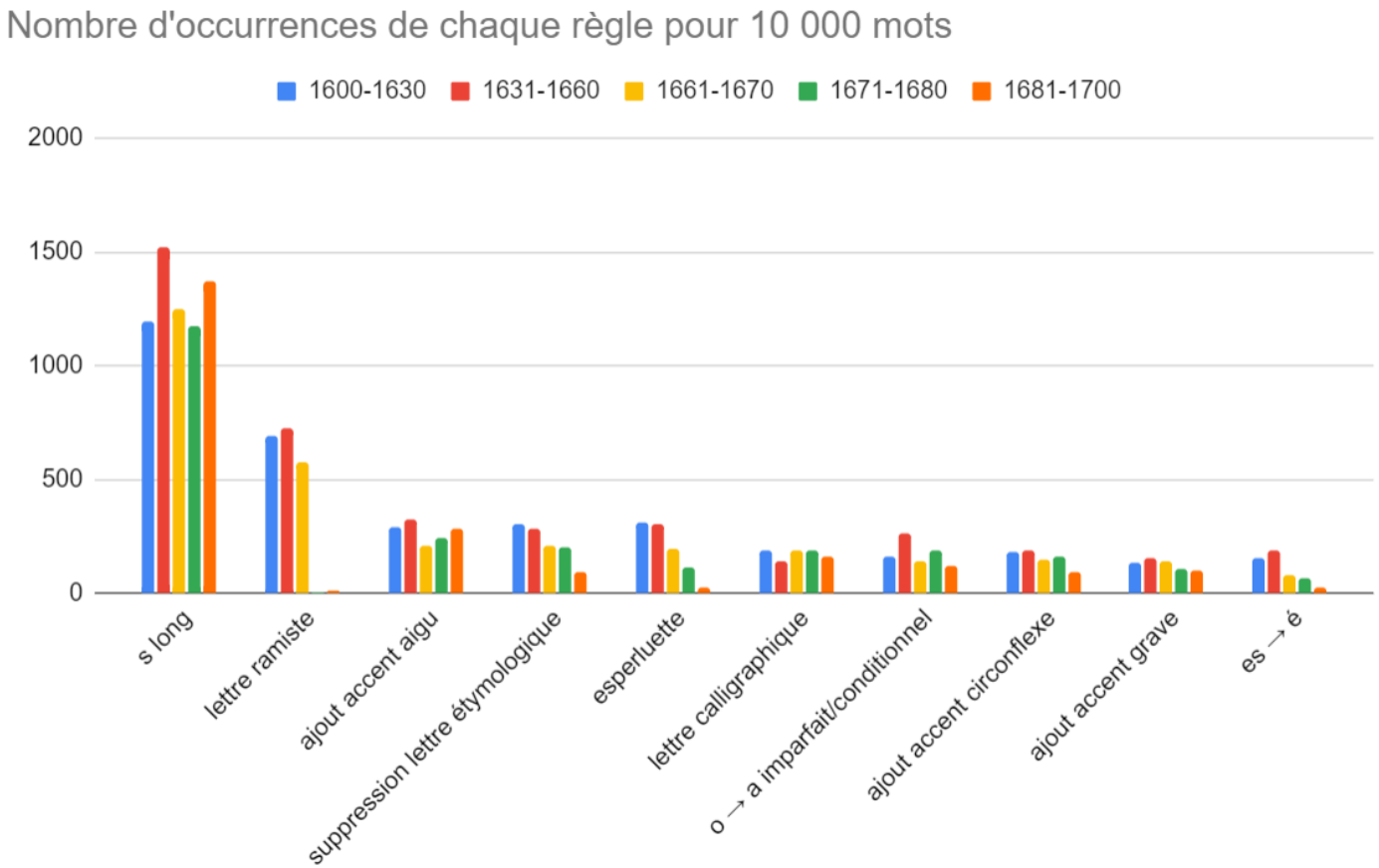
Les transformations (substitution, suppression, insertion) sont regroupées selon des grands types en s'appuyant sur la littérature existante (par ex. Vachon 2010).

Exemples :

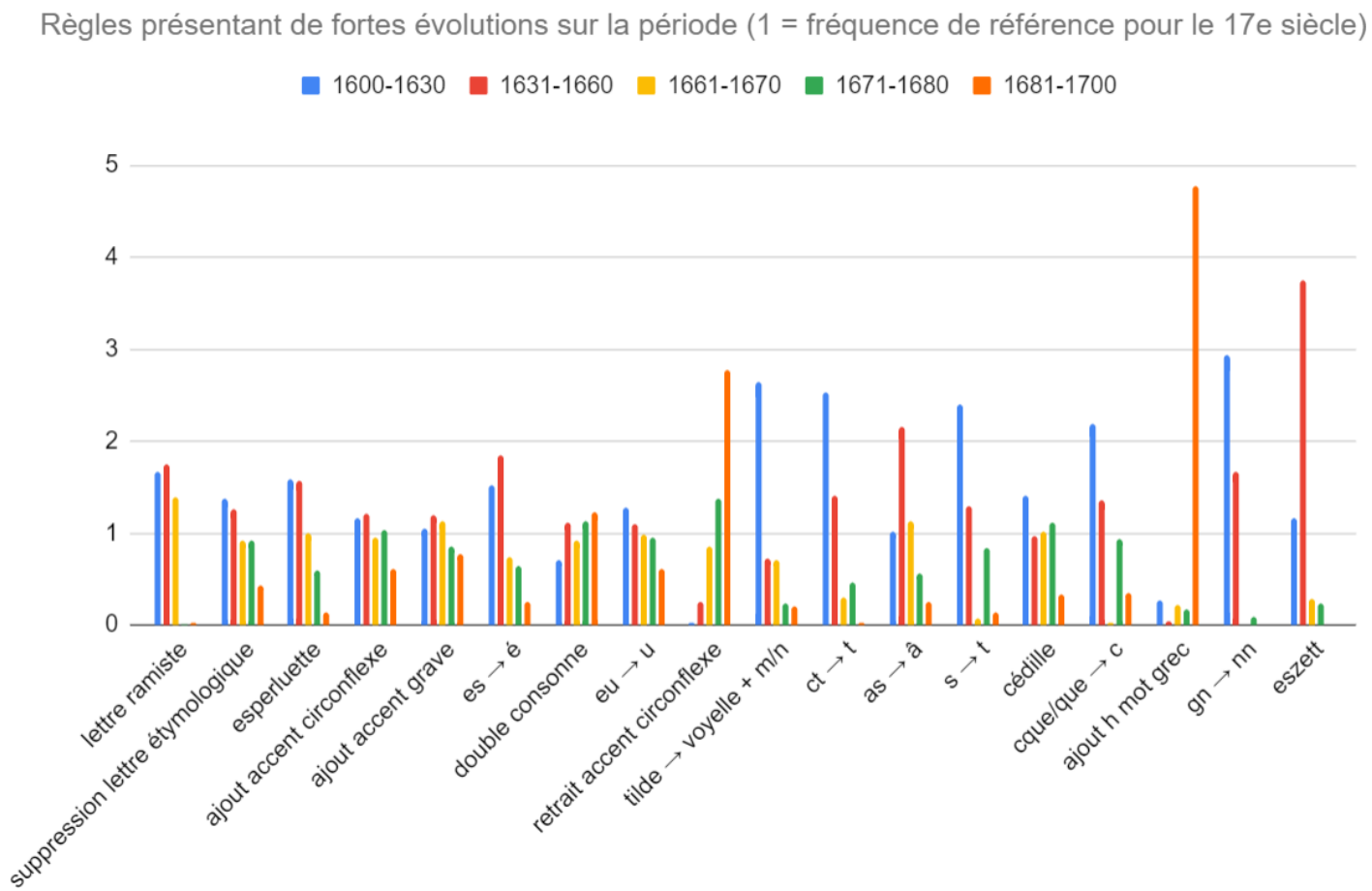
- lettre ramiste : un *i*, un *j*, un *u*, un *v* dans le mot en version originale associé respectivement à un *j*, un *i*, un *v*, un *u* en version normalisée (*vniuers*→*univers*) ;
- lettre calligraphique : un *y* en fin de mot en version originale est associé à un *i* en version normalisée (*roy*→*roi*) ;
- lettre diacritique : *s* dans *hospital estoit* remplacé par un accent circonflexe (*hôpital*) ou aigu (*était*).

[Claire Hélène Vachon, *Le Changement linguistique au XVI^e s. : une étude basées ur des textes littéraires français*, Strasbourg, France, 2010]

Résultats sur le corpus découpé en 5 périodes



Évolutions significatives au cours du 17e siècle

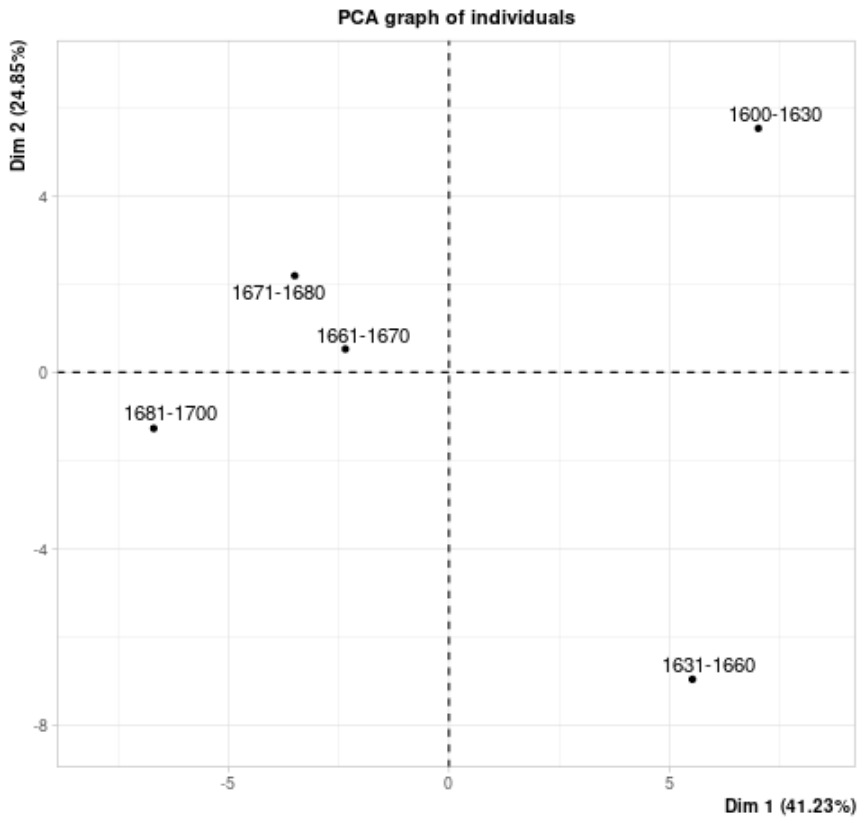


Une valeur de 1 correspond à la valeur moyenne de la fréquence de la règle (nombre d'occurrences divisé par nombre total de mots de l'ensemble du corpus).

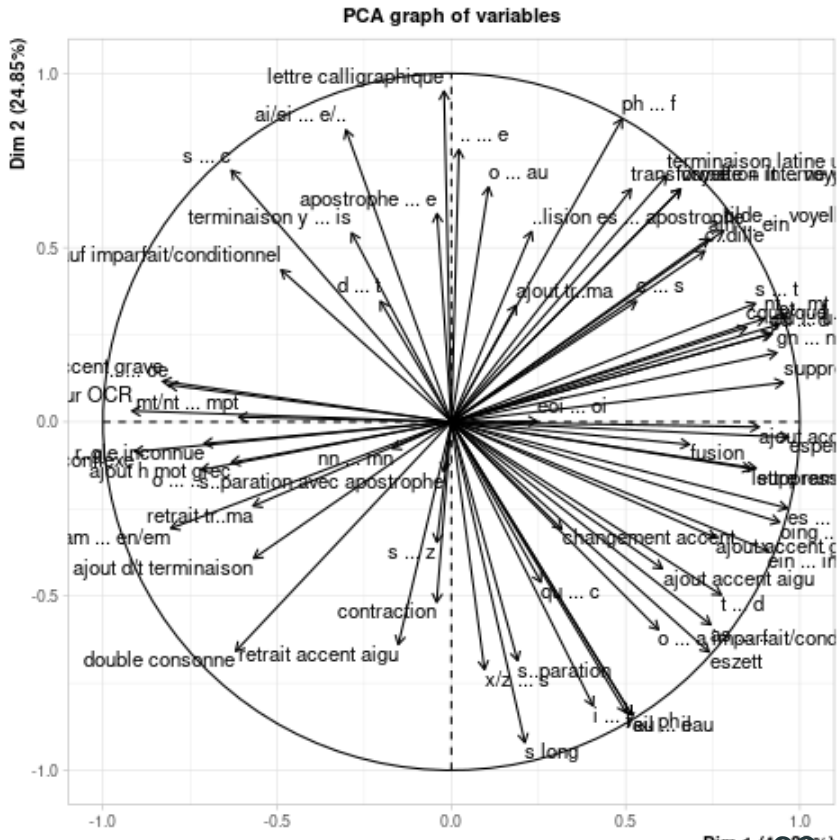
Au-delà de la diachronie grâce aux métadonnées

Quels paramètres dans les métadonnées des textes (date, lieu, genre, auteur…) permettent de donner un sens à ces règles ? Une analyse en composantes principales organise les résultats en deux axes, qui sont potentiellement deux critères de classification.

(a. graphique des périodes)



(b. graphique des règles)



Plan de la présentation

Des *scriptae* en français classique

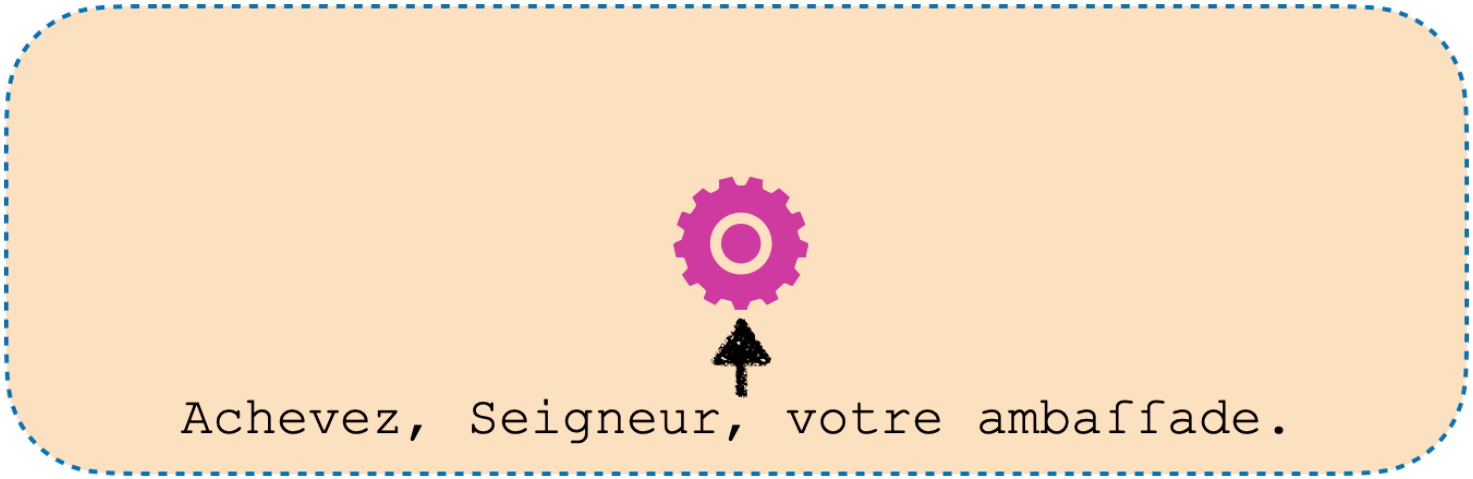
De l'utilité de la scriptométrie

Approche par règles

Approche par réseau de neurones

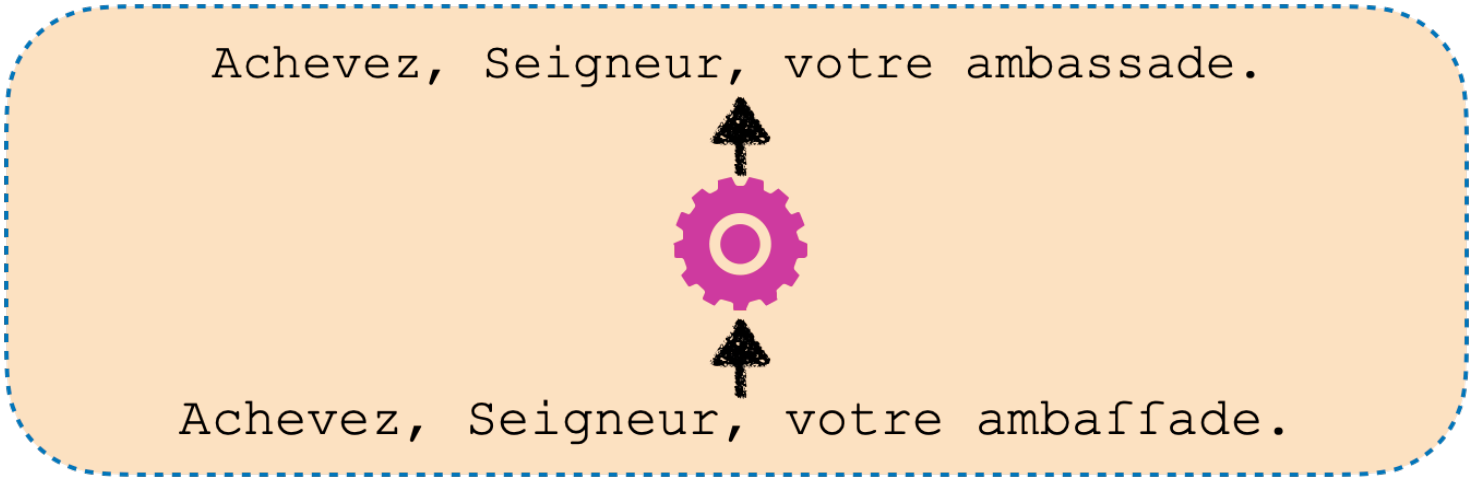
Conclusion

Approche par réseau de neurones



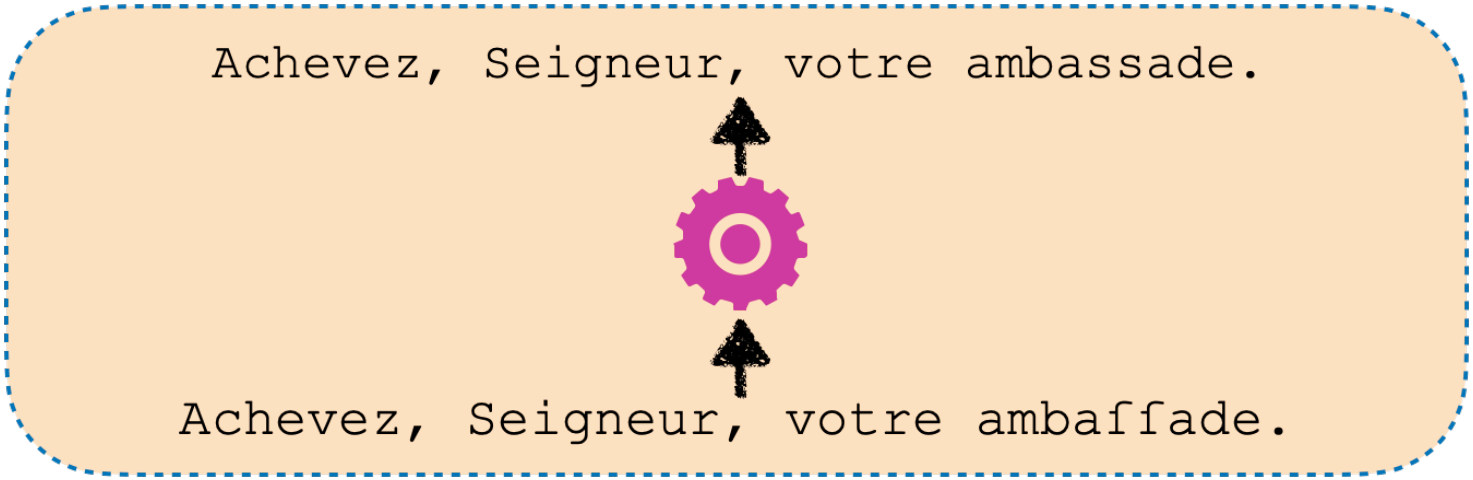
- Un modèle de normalisation automatique (un modèle à base de réseau de neurones) entraîné sur PARALLEL17

Approche par réseau de neurones



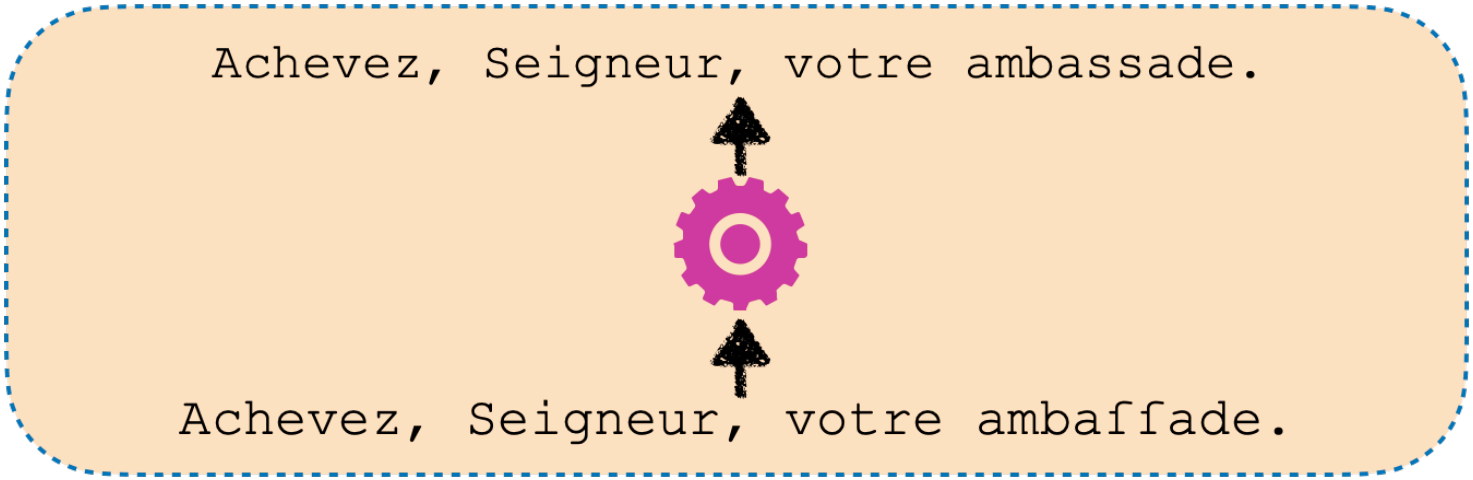
- Un modèle de normalisation automatique (un modèle à base de réseau de neurones) entraîné sur PARALLEL17

Approche par réseau de neurones



- Un modèle de normalisation automatique (un modèle à base de réseau de neurones) entraîné sur PARALLEL17
- Que pouvons-nous apprendre d'un tel modèle de normalisation, concernant la variation graphique ?

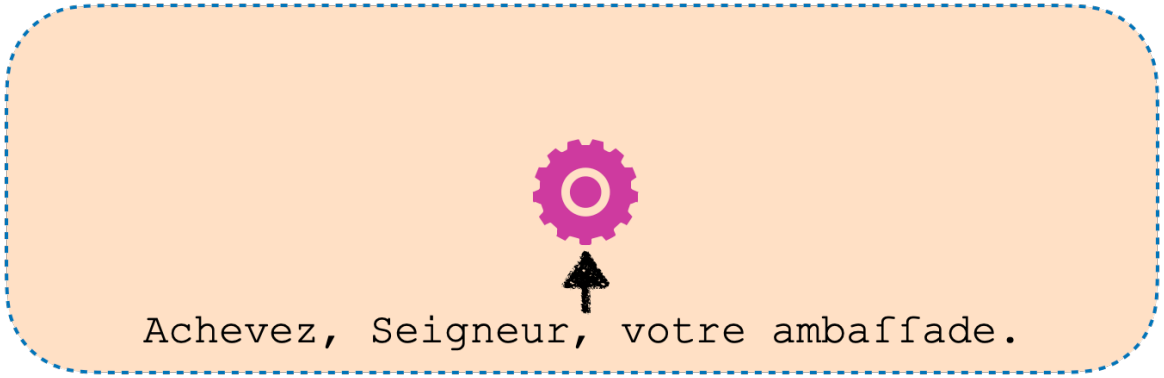
Approche par réseau de neurones



- Un modèle de normalisation automatique (un modèle à base de réseau de neurones) entraîné sur PARALLEL17
- Que pouvons-nous apprendre d'un tel modèle de normalisation, concernant la variation graphique ?
- Deux expériences préliminaires :
  1. Le modèle est-il sensible au changement linguistique ?
  2. Comment extraire les informations apprises par le modèle ?

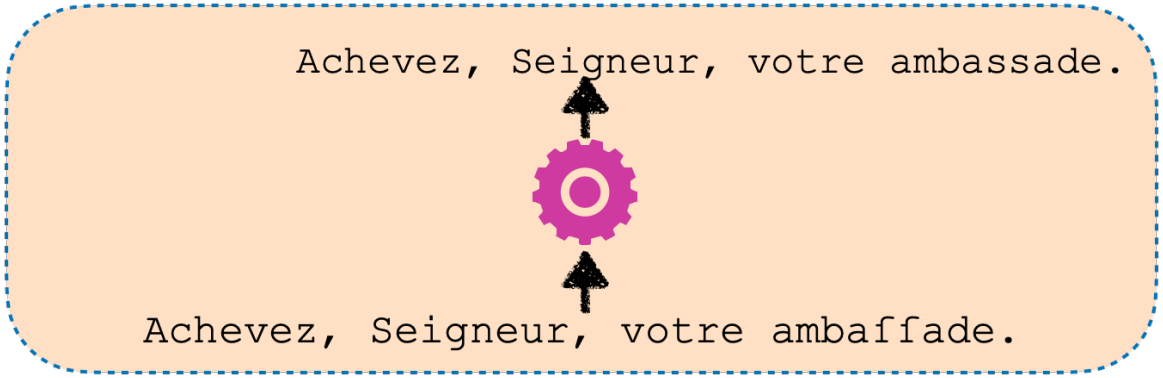
1. Sensibilité au changement linguistique

- Y a-t-il suffisamment d'informations dans les phrases originales pour prédire le texte normalisé ***et* la décennie de rédaction** ?



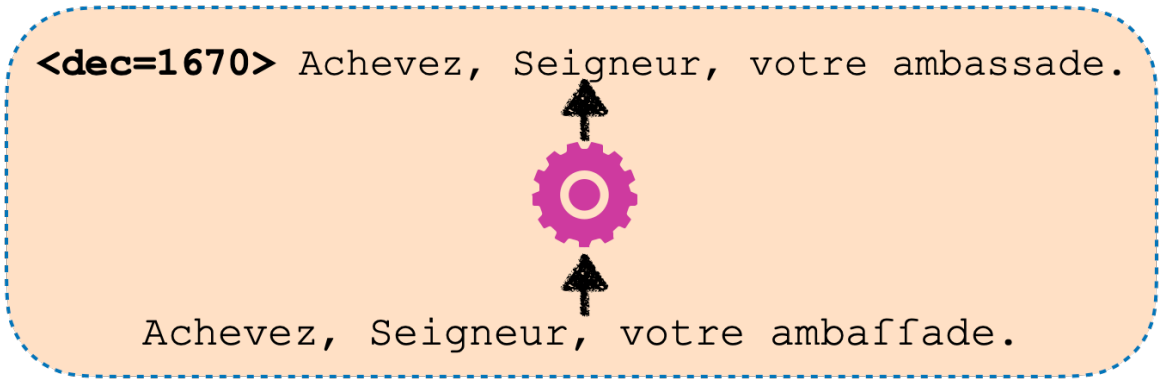
1. Sensibilité au changement linguistique

- Y a-t-il suffisamment d'informations dans les phrases originales pour prédire le texte normalisé ***et* la décennie de rédaction** ?



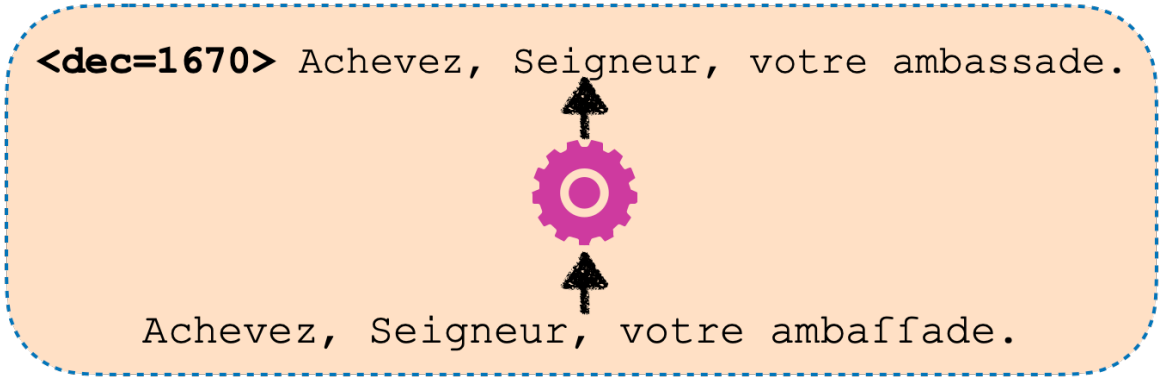
1. Sensibilité au changement linguistique

- Y a-t-il suffisamment d'informations dans les phrases originales pour prédire le texte normalisé ***et* la décennie de rédaction** ?



1. Sensibilité au changement linguistique

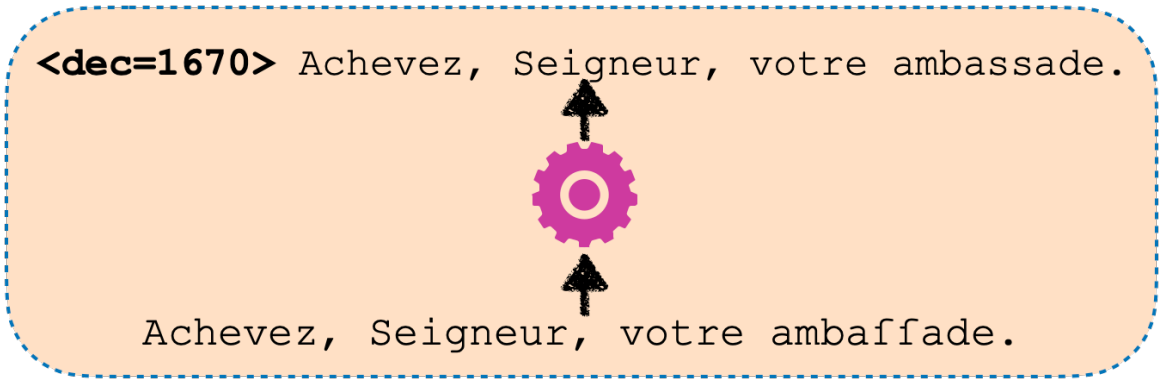
- Y a-t-il suffisamment d'informations dans les phrases originales pour prédire le texte normalisé ***et* la décennie de rédaction** ?



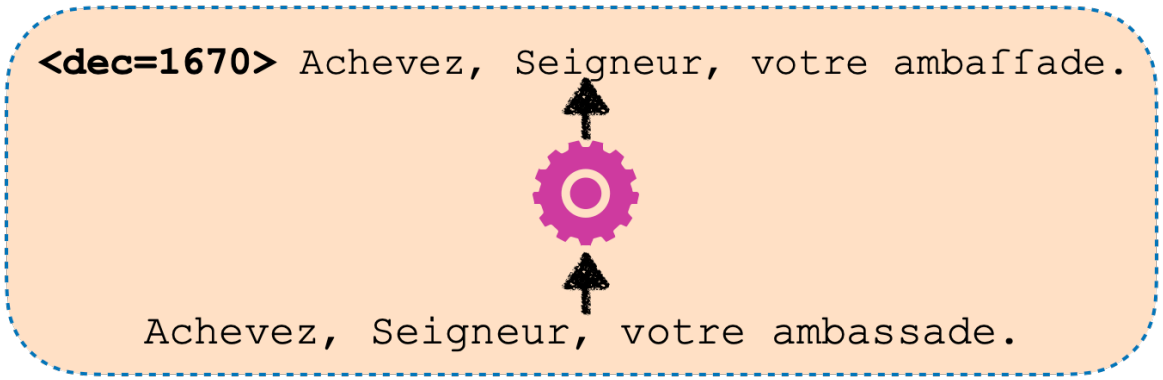
- Mais les informations apprises sont-elles de nature graphique ou lexicale ?

1. Sensibilité au changement linguistique

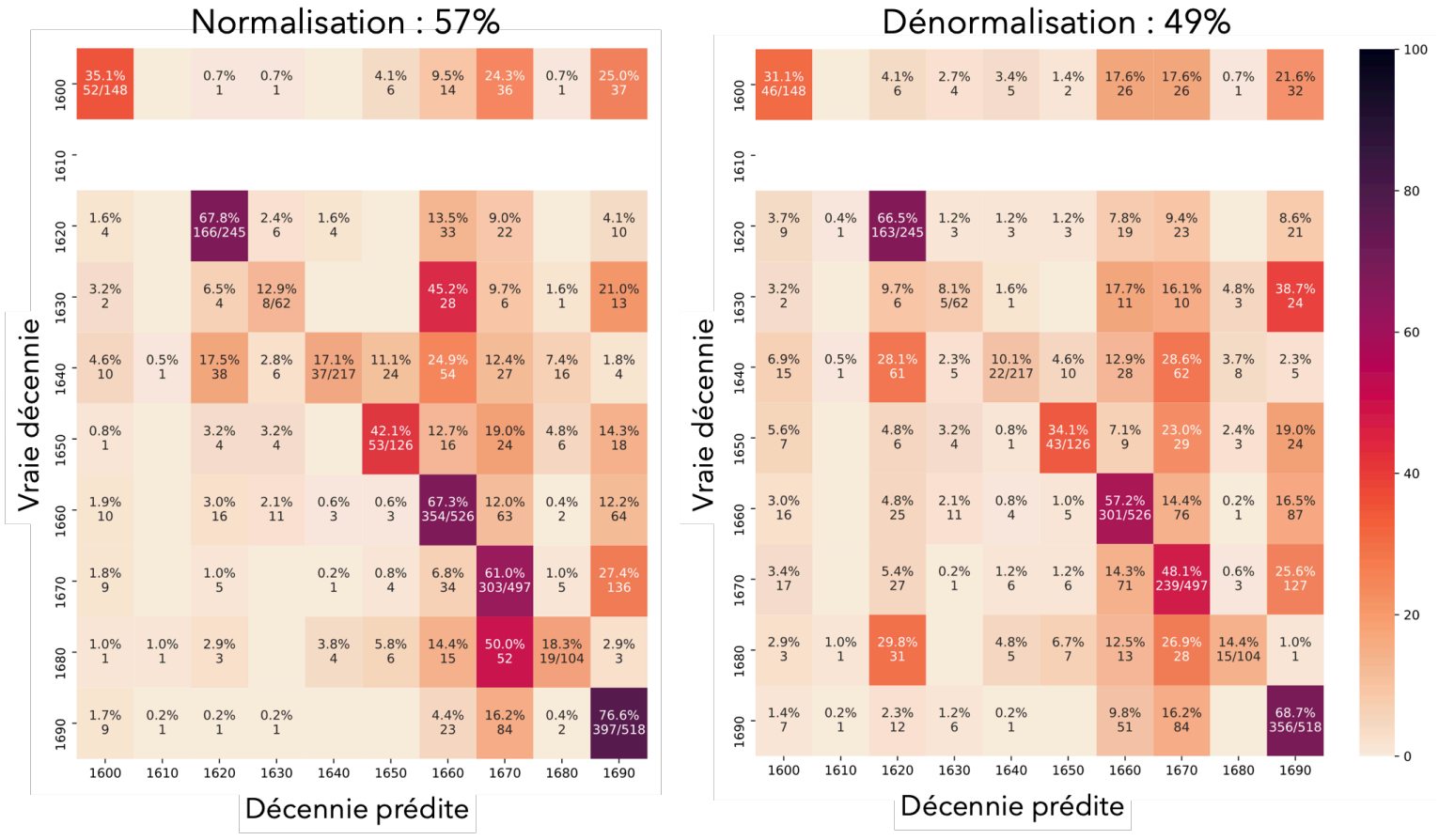
- Y a-t-il suffisamment d'informations dans les phrases originales pour prédire le texte normalisé ***et* la décennie de rédaction** ?



- Mais les informations apprises sont-elles de nature graphique ou lexicale ? **Expérience de contrôle : dénormalisation (réduction de l'information au simple lexique)**

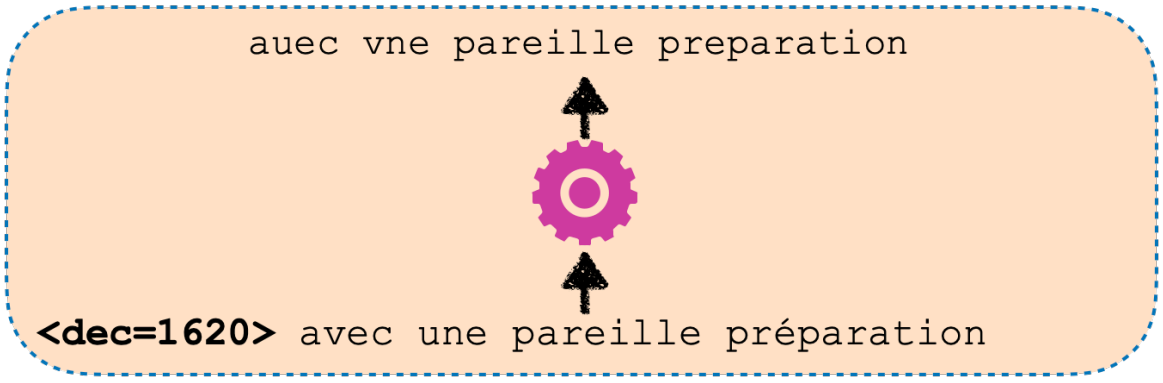


1. Sensibilité au changement linguistique - résultats



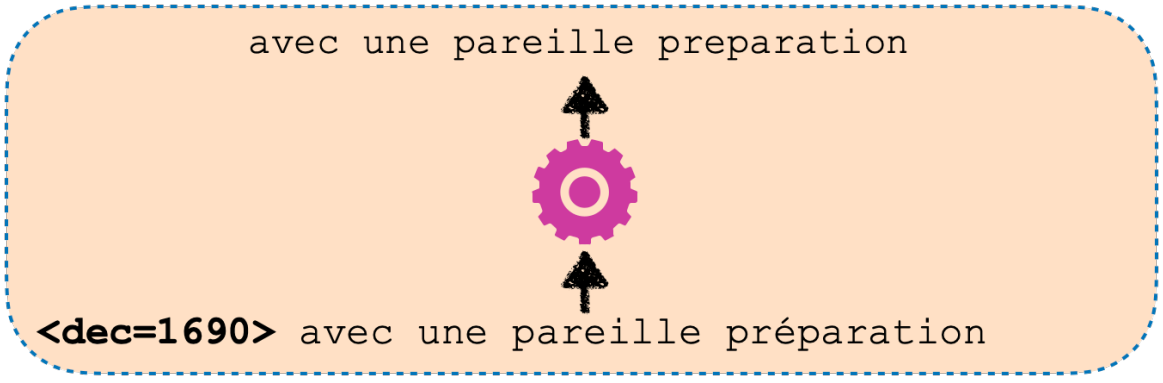
- Prédictions assez bonnes pour les deux modèles
- Les scores sont plus élevés quand le modèle a accès à la variation graphique
- Le modèle ne s'appuie donc pas uniquement sur le lexique pour prédire la décennie

2. Quelles informations apprises ?



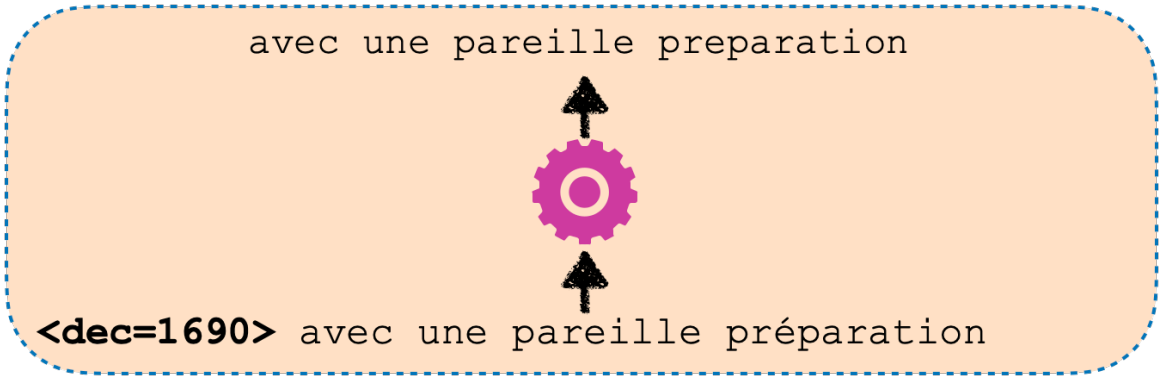
- Modèle de dénormalisation, conditionné sur la décennie

2. Quelles informations apprises ?



- Modèle de dénormalisation, conditionné sur la décennie
- On peut donc créer une version « classique » artificielle pour chaque décennie (1600, 1610, 1620 ...)

2. Quelles informations apprises ?



- Modèle de dénormalisation, conditionné sur la décennie
- On peut donc créer une version « classique » artificielle pour chaque décennie (1600, 1610, 1620 ...)
- L'objectif est de concevoir un corpus parfaitement homogène, où chaque décennie est parfaitement comparable à une autre.

2. Quelles informations apprises ? - exemples

Phrase normalisée (en entrée) :
Un soufflet bien fermé de tous **côtés** fait le **même** effet, **avec une** pareille **préparation**

Phrase denormalisée, conditionnée sur 1620 :
Vn souflet bien fermé de tous **costez** fait le **mesme** effet, **auec vne** pareille **preparation**

Phrase denormalisée, conditionnée sur 1690 :
Vn souflet bien fermé de tous **costez** fait le **mesme** effet, **avec une** pareille **preparation**

- Créer des versions « classiques » pour des textes normalisés du corpus D'ALEMBERT
- Aligner le résultat pour chaque décennie avec le texte normalisé et extraire les règles qui permettent de passer d'une version à l'autre, par exemple *une*>*vne*
- Comparer les différentes règles trouvées pour chaque décennie et tenter de dégager une évolution de celles-ci

Quelles informations apprises ? - résultats

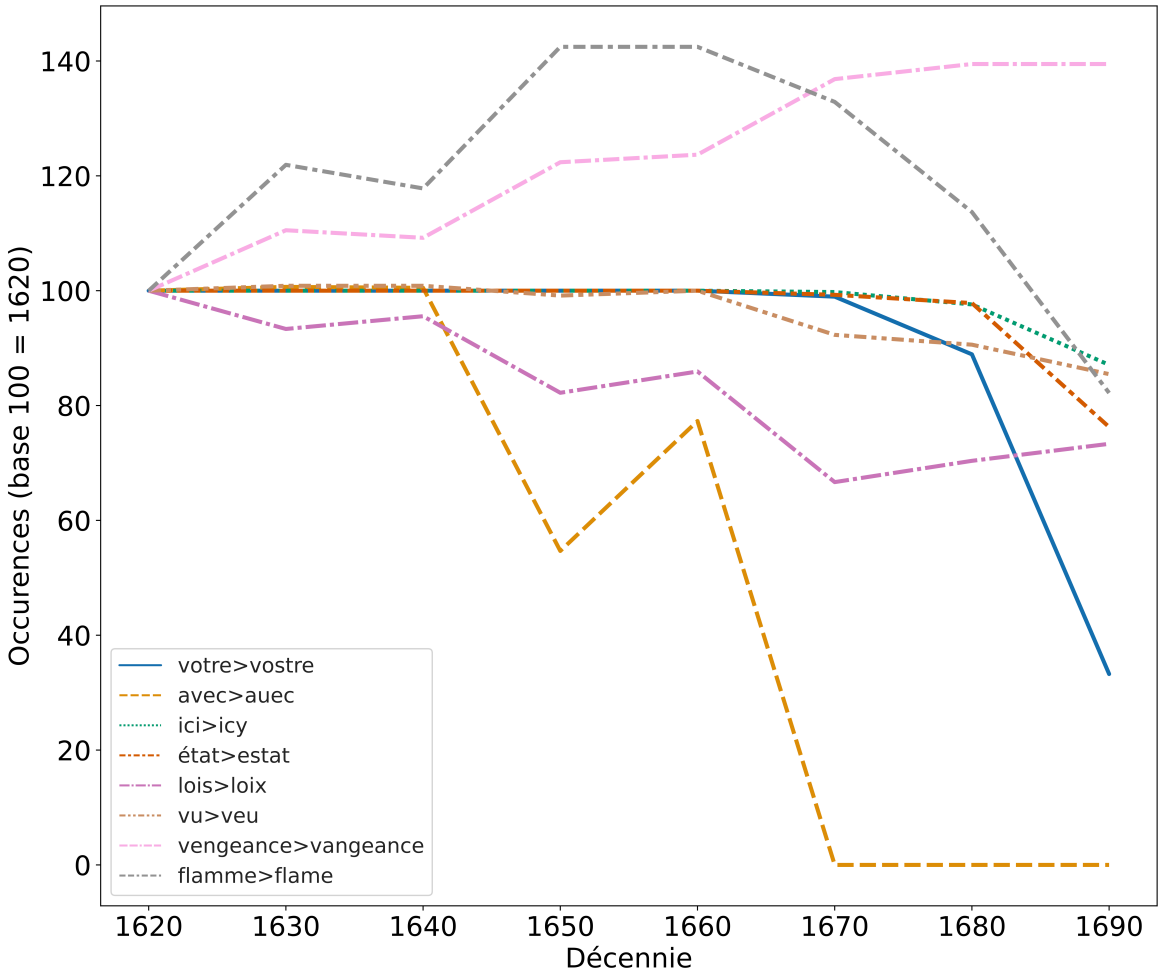


Figure 2 – Evolution de plusieurs règles de dénormalisation appliquées sur les textes normalisés du corpus D'ALEMBERT par périodes de 10 ans entre 1620 (la base 100) et 1700.

Plan de la présentation

Des *scriptae* en français classique

De l'utilité de la scriptométrie

Approche par règles

Approche par réseau de neurones

Conclusion

Conclusion

Conclusions préliminaires

Nous sommes en mesure de tirer des premières conclusions

- Nous sommes capable de réduire la fenêtre chronologique des analyses à une décennie (contre une vingtaine d'années pour les études de Cl. Vachon)
- Contrairement aux époques précédentes, la période classique ne semble pas témoigner de phase de reflux
- Nous pouvons postuler l'existence d'une nouvelle phase graphématique autour des années 1670.

Problèmes méthodologiques

Existence d'un possible angle mort du fait de la méthode même

- La *scripta* est évaluée à l'aune de sa distance avec le français contemporain
- Idéal pour étudier la stabilisation du système graphique
- Méthode valable pour étudier la couleur d'un texte entre deux pôles ?
- Comment analyser les évolutions de *tempus*, repéré comme moderne si graphié *tems* mais pas comme ancien si graphié *temps* ?

Further work

Il reste de multiples pistes à explorer

- Réintroduire le « pallier textuel », essentiel en philologie, pour caractériser précisément la nature des textes
- Intégrer les manuscrits dans le corpus, moins sujets que les imprimés à la neutralisation graphématique et à la coexistence de différents systèmes graphiques en diasystème
- Reprendre l'hypothèse de l'existence d'une *scripta* « habituelle », dont il conviendrait de définir plus précisément la nature de son hétérogénéité
- Reproduction de la méthode sur des données marquées par des dialectes

Remerciements

Alexandre Bartz, Pedro Ortiz-Suarez... et Myriam pour l'invitation !