Méthodes pour l'archéologie linguistique : datation par combinaison d'indices temporels

Anne Garcia-Fernandez Anne-Laure Ligozat
Marco Dinarelli Delphine Bernhard
LIMSI-CNRS

DEFT 2011, Montpellier, France

Plan

Exemple de document

La séance musicale de M. Félicien David au Palais de l'Industrie a obtenu un succès complet les fragmens du Désert, de Christophe Colomb et de Moïse au Sinaï ont été très vivemçnt applaudis ; le Chant du soir a été redemandé par une acclamation unanime. Jeudi 22, le même programme sera de nouveau exécuté dans les mêmes conditions : l,2S0 choristes et'instrumentistes. Samedi 24, seconde exécution du concert dirigé par M. Berlioz. Dimanche 2S, fermeture de la nef centrale du Palais de l'Industrie et clôture des fêtes musicales. Lotecfètairedela rédaction, F. Carani.

Indices temporels

- Date postérieure à la naissance des personnes citées
- Termes archaïquement orthographiés
- Archaïsmes et néologismes

Exemple de document

La séance musicale de M. Félicien David au Palais de l'Industrie a obtenu un succès complet les fragmens du Désert, de Christophe Colomb et de Moïse au Sinaï ont été très vivemçnt applaudis ; le Chant du soir a été redemandé par une acclamation unanime. Jeudi 22, le même programme sera de nouveau exécuté dans les mêmes conditions : l,2S0 choristes et'instrumentistes. Samedi 24, seconde exécution du concert dirigé par M. Berlioz. Dimanche 2S, fermeture de la nef centrale du Palais de l'Industrie et clôture des fêtes musicales. Lotecfètairedela rédaction, F. Carani.

Indices temporels

- Date postérieure à la naissance des personnes citées
- Termes archaïquement orthographiés
- Archaïsmes et néologismes

Exemple de document

La séance musicale de M. Félicien David au Palais de l'Industrie a obtenu un succès complet les fragmens du Désert, de Christophe Colomb et de Moïse au Sinaï ont été très vivemçnt applaudis ; le Chant du soir a été redemandé par une acclamation unanime. Jeudi 22, le même programme sera de nouveau exécuté dans les mêmes conditions : l,2S0 choristes et'instrumentistes. Samedi 24, seconde exécution du concert dirigé par M. Berlioz. Dimanche 2S, fermeture de la nef centrale du Palais de l'Industrie et clôture des fêtes musicales. Lotecfètairedela rédaction, F. Carani.

Indices temporels

- Date postérieure à la naissance des personnes citées
- Termes archaïquement orthographiés
- Archaïsmes et néologismes

Description générale

Méthodes chronologiques

- À partir des indices temporels

Similarité temporelle

- Calcul de similarité cosinus entre portions
- Utilisation d'un SVM

Généralités

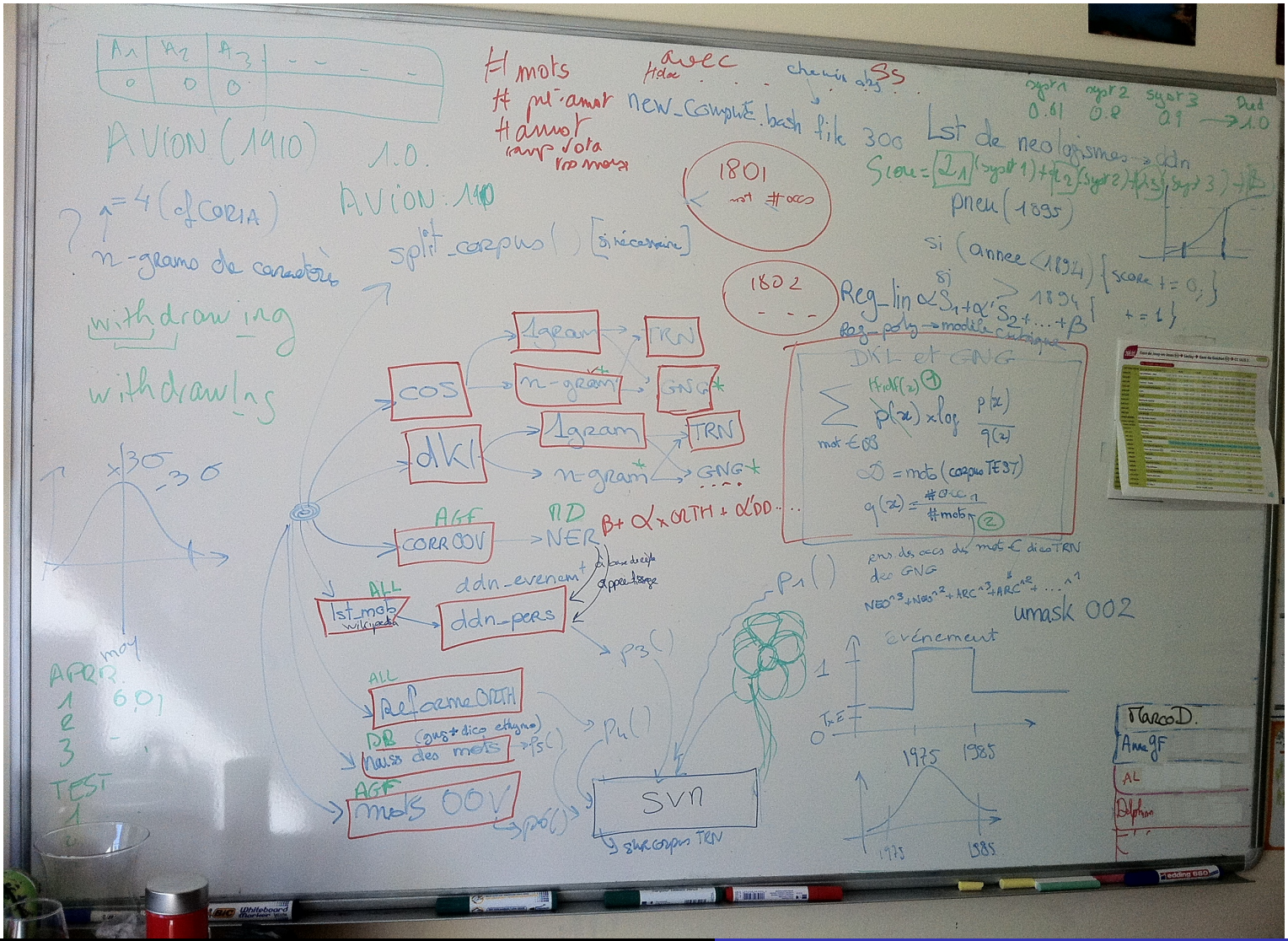
Prétraitements

- Corpus divisé en deux : TRN (2396 portions) et DEV (1200)
- Lemmatisation (TreeTagger)

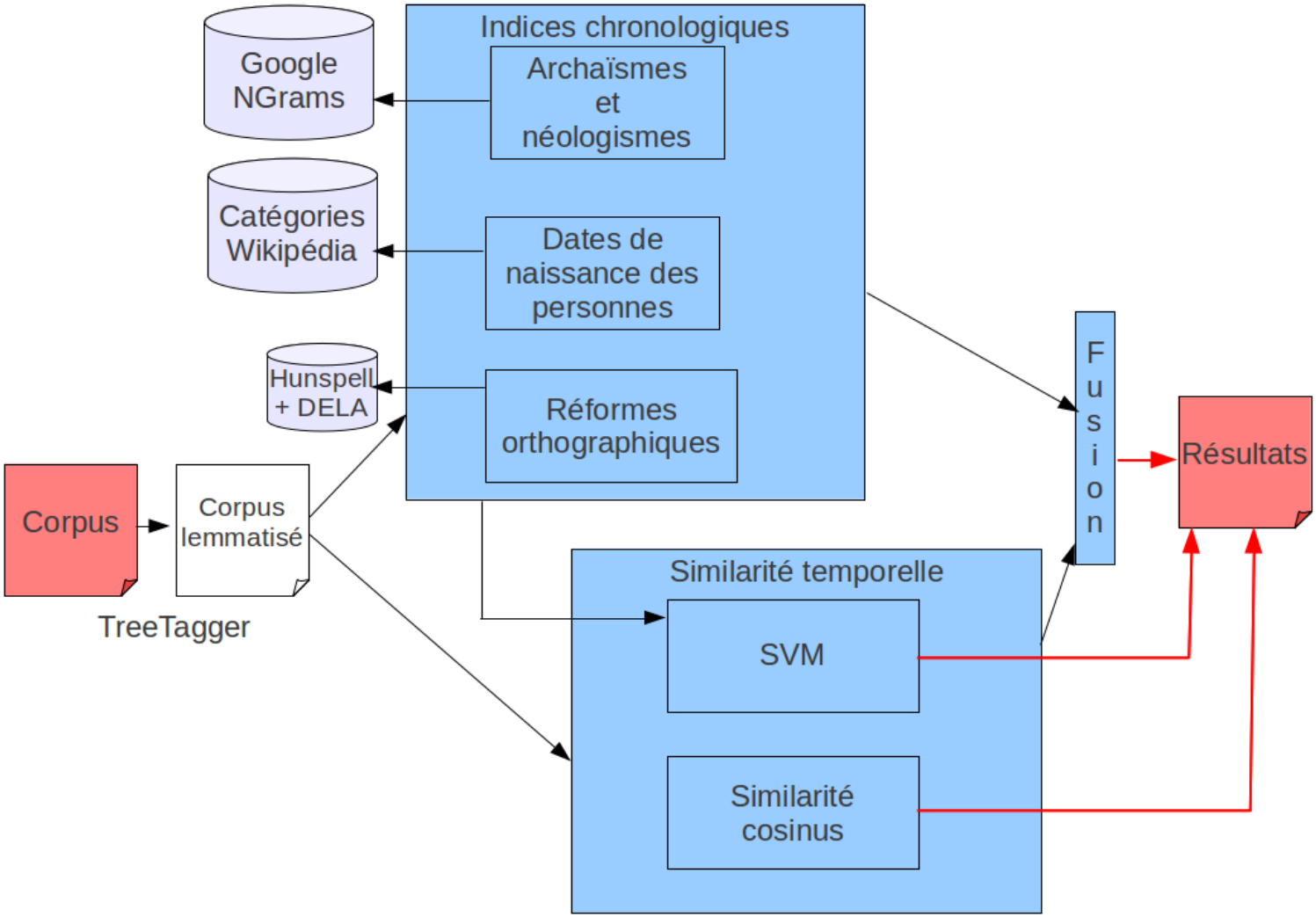
Ressources et outils utilisés

- Wikipédia
- Google Books Ngrams
- Hunspell et le dictionnaire DELA

Architecture du système ☺



Architecture du système



Plan

Dates de naissance de personnes 1/3

Méthode

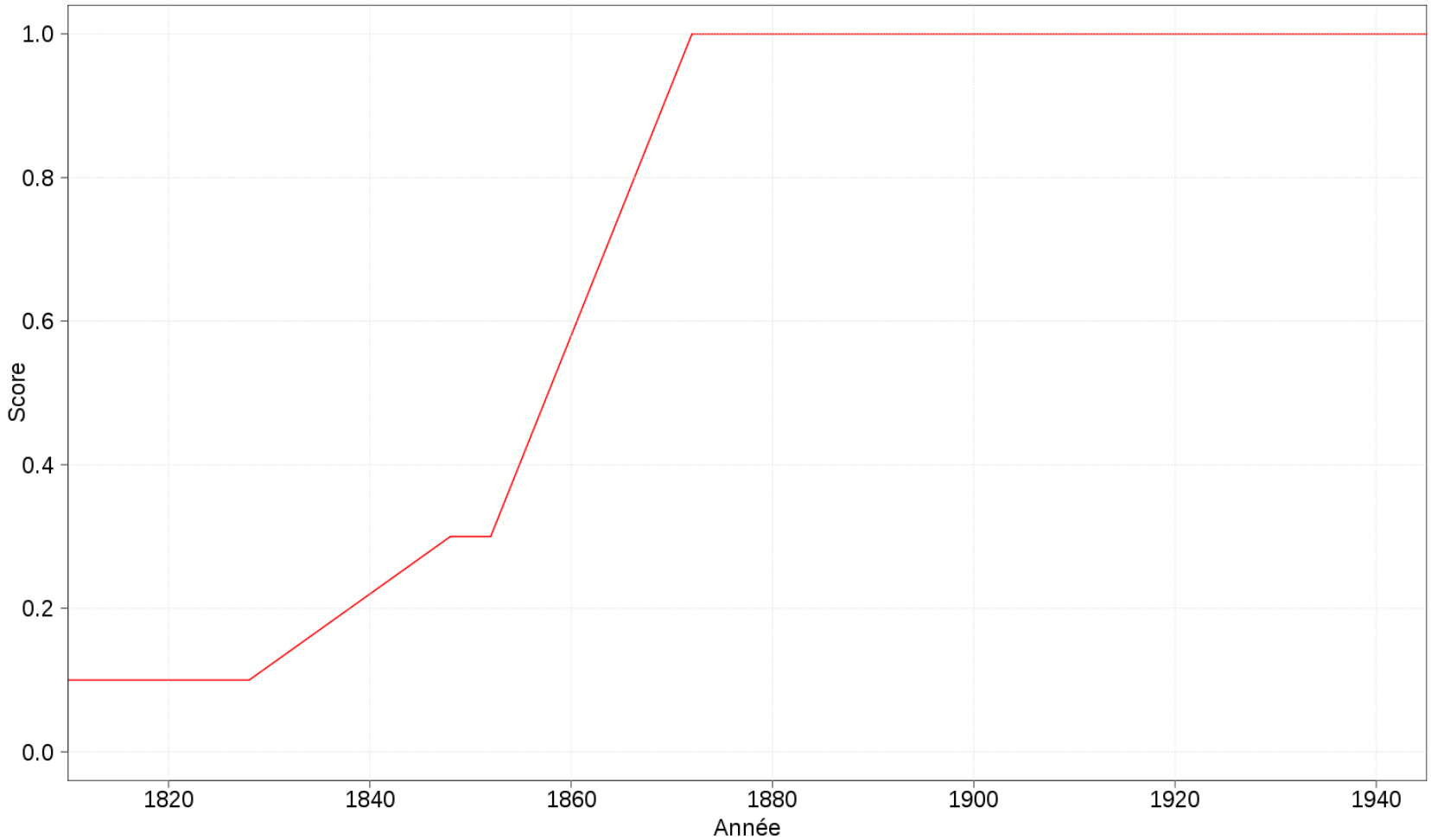
- Utilisation des pages Catégorie :Naissance_en_AAAA de la Wikipédia
- Détection des noms de personnes au sein des portions
- Attribution d'une probabilité pour chaque portion pour chaque année

Exemple d'indice

Parlant plus loin des lois récemment votées et en particulier des assurances sociales , M . Louis Barthou affirme à ses lecteurs que « le devoir parlementaire n'est pas toujours commode »
$\rightarrow$ né en 1862

Dates de naissance de personnes 2/3

avec Jules Verne, né en 1828 et Antoni Gaudí, né en 1852



Dates de naissance de personnes 3/3

Évaluation

- Ressource :
  - téléchargement des pages pour les 144 années
  - 96 000 noms extraits
- Annotation du corpus TRN :
  - 529 noms de personnes
  - 375 portions (16% des portions)
- Qualité de l'indice :
  - 3% d'erreurs (homonymes, annotations erronées)

- Mlle Colette Burin des Roziers dont le mariage avec le baron Honoré de Rascas de Châteauredon vient d'être béni en l'église SaintFrançois-Xavier.

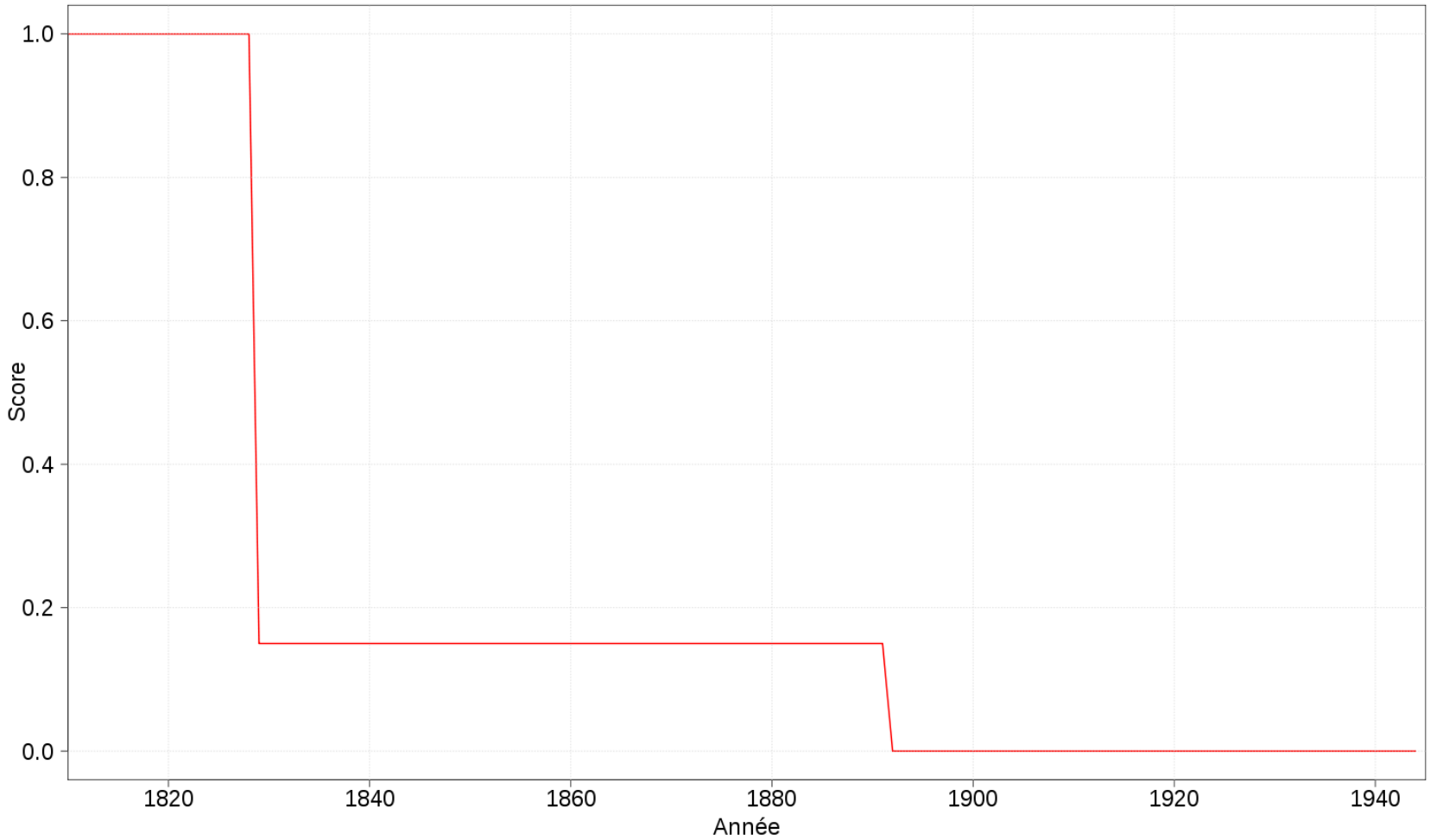
Réformes orthographiques 1/3

Méthode

- Deux réformes du français entre 1801 et 1944
- Identification des termes orthographiés selon les règles orthographiques valables :
  - avant la réforme de 1835 (w_{ref1}) :
    - mots inconnus en "ois/oit/oient" dont la forme équivalente en "ais/ait/aient" existe
  - avant la réforme de 1878 (w_{ref2})
    - mots inconnus en "ans/ens" dont la forme équivalente en "ants/ents" existe
- Adaptation au corpus (sur TRN)
  - w_{ref1} sont majoritairement présents avant **1828**
  - w_{ref2} sont présents avant **1891**

Réformes orthographiques 2/3

avec *appartemens* (réforme 1891) et *faudroit* (réforme 1828)



Réformes orthographiques 3/3

Évaluation

- Détection dans le corpus TRN :
  - 864 termes considérés comme antérieurs à 1828
  - 367 termes considérés comme antérieurs à 1891
  - 655 portions concernées (27% des portions)
- Qualité de l'indice :
  - 0,5% d'erreurs pour la réforme de 1828, 0 pour 1891
  - causé par des erreurs d'OCRisation :

- *une tois pour toutes*

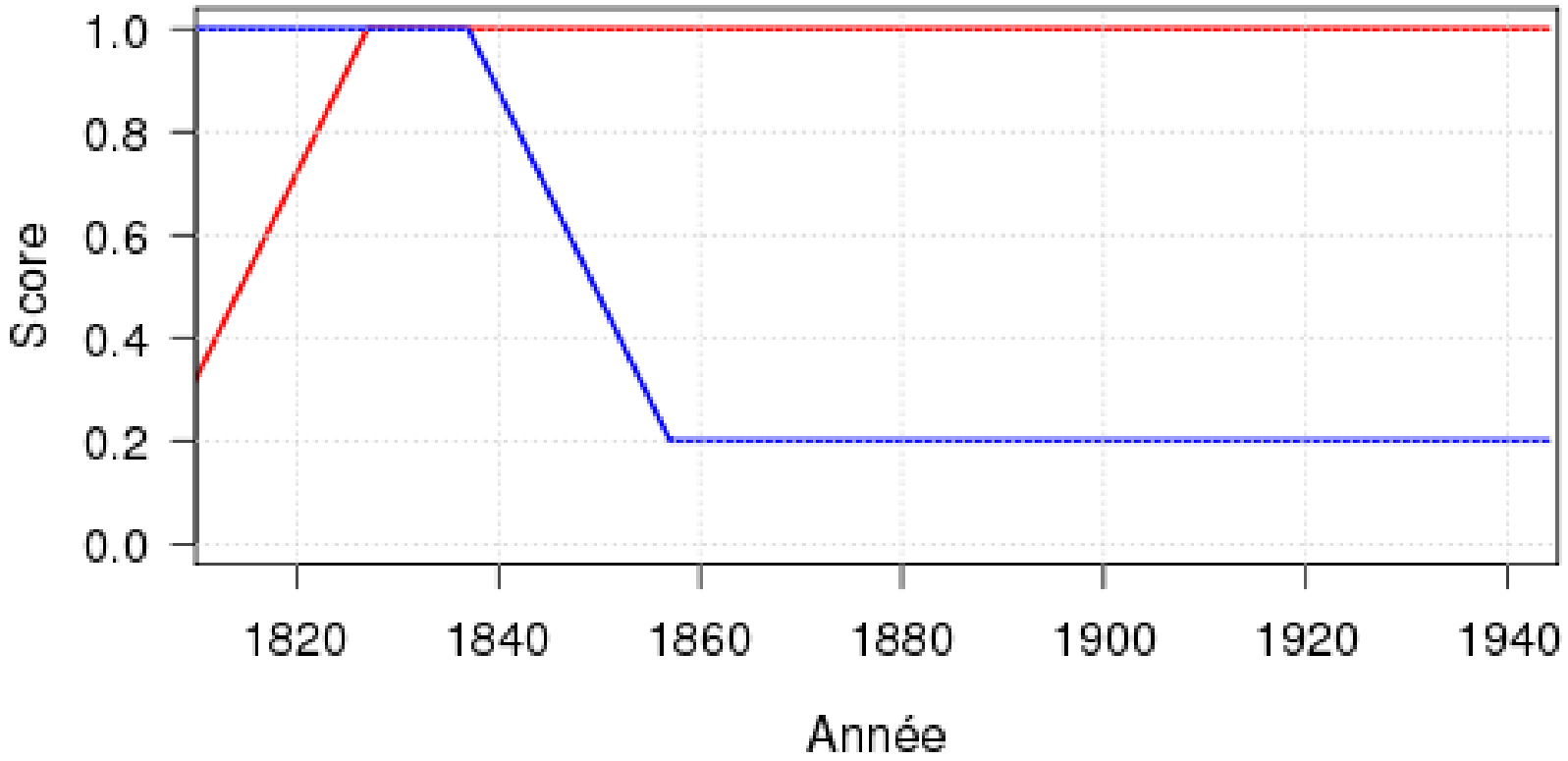
Néologismes et archaïsmes 1/3

Méthode

- Recherche des dates d'apparition et de disparition de termes
- Fréquence d'occurrences de termes issus des Google Books Ngrams
- Détermination de seuil de fréquence pour déterminer la date d'apparition et/ou de disparition d'un terme

Néologismes et archaïsmes 2/3

avec *Colombie* (néologisme daté de 1827) et amans (archaïsme daté de 1837)



Néologismes et archaïsmes 3/3

Évaluation

- Ressource acquise :
  - 114 396 néologismes
  - 53 392 archaïsmes
- Détection du corpus TRN :
  - 175 002 néologismes (dont 30 233 pour 1801...), 34 202 archaïsmes
  - 2396 portions concernées (100% des portions)
- Qualité de l'indice :
  - 3% d'erreurs pour les néologismes et 10% pour les archaïsmes
    - très fiable mais relativement peu précis

Plan

Similarité cosinus

Méthode

- Calcul de la similarité cosinus entre :
  - une nouvelle portion
  - et des données de référence pour une année donnée
- *Rappel : mesure de similarité entre deux vecteurs*
- Corpus de référence :
  - corpus d'entraînement : portions regroupées par année
  - Google Ngrams
- Vecteurs :
  - de tf.idf (adaptés)
  - mots et caractères (car données bruitées)
  - utilisation des n-grams (n allant de 1 à 6)

SVM 1/3

Méthode

- Utilisation de `svm-light`
- Noyau polynomial
- Approche *un-contre-tous* (car peu de données) :
  144 modèles binaires, un pour chaque année
- Évaluation : chaque exemple de test est classifié par chaque modèle

SVM 2/3

Système DEFT 2011

- n-grams de mots et lemmes, $n = 2$
- Pondération des traits f : # d'occurrences
- Intégration des informations chronologiques :
  - dates de naissance, néologismes, archaïsmes, orthographe
  - deux traits par indices : présent/absent et présence d'un néologisme pour une année

SVM 3/3

Dernier système

- n-grams de caractères de lemmes, $n = 6$
- Pondération des traits : $tf \cdot idf$
- Intégration des informations chronologiques
  - avec une pondération plus importante
- Normalisation des vecteurs : $\vec{X} = (\frac{x_1}{||X||}, ..., \frac{x_m}{||X||})$
- Fenêtre w : les instances positives sont les années dans une fenêtre w autour de l'année cible avec w=5 pour les portions de 300 et w=3 pour les 500

Plan

Sélection de la meilleure année

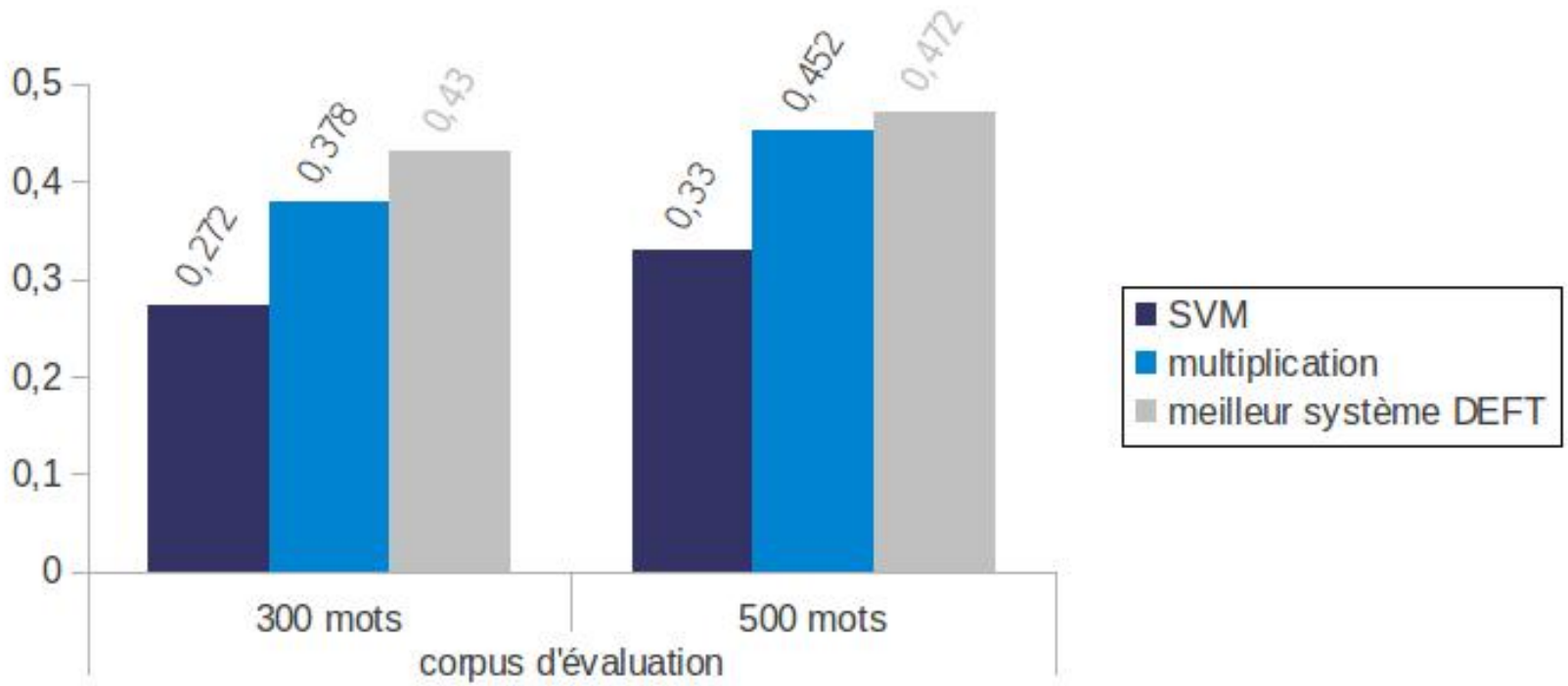
Méthode

- Multiplication des scores
- Sélection de l'année ayant le meilleur score

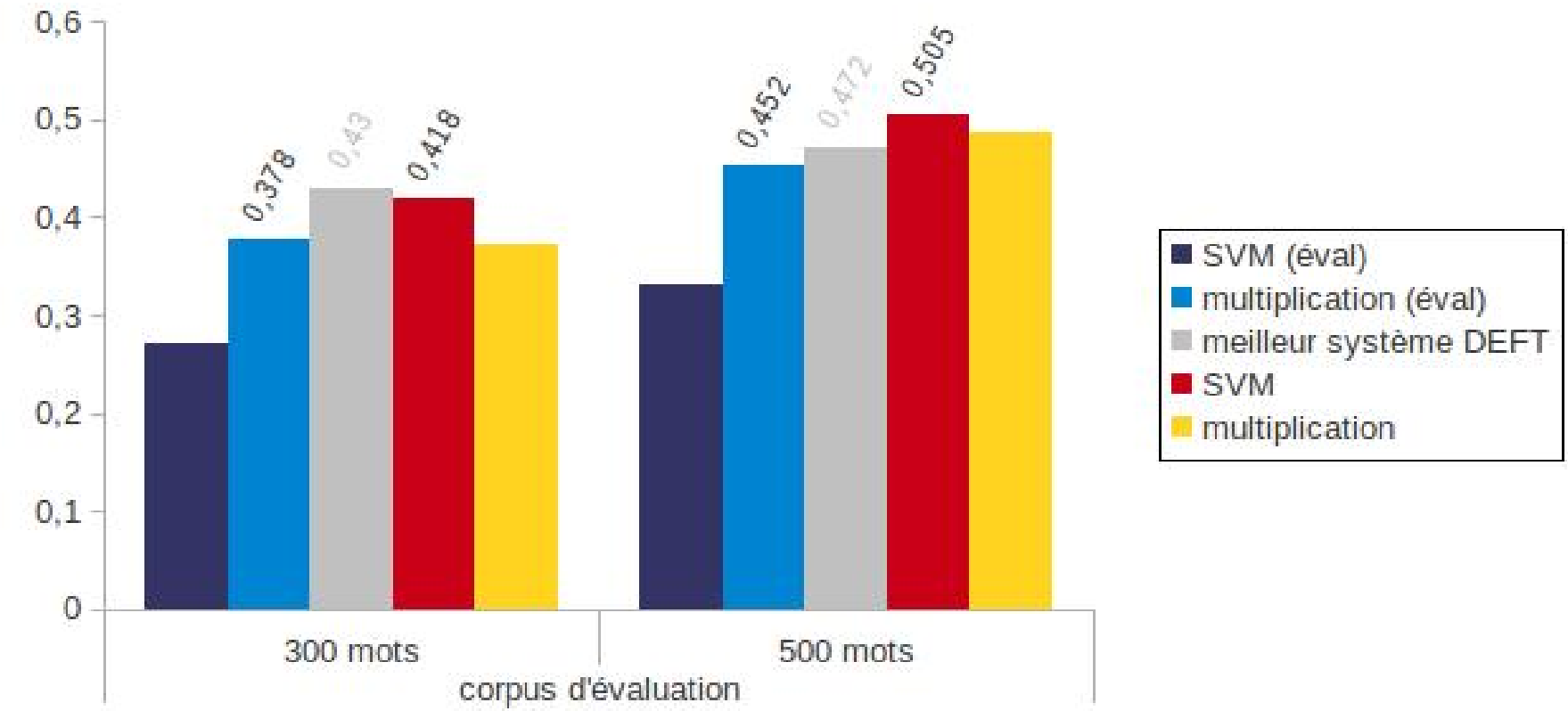
Évaluation

- Mesure d'évaluation DEFT 2011
- % d'années et de décennies correctes

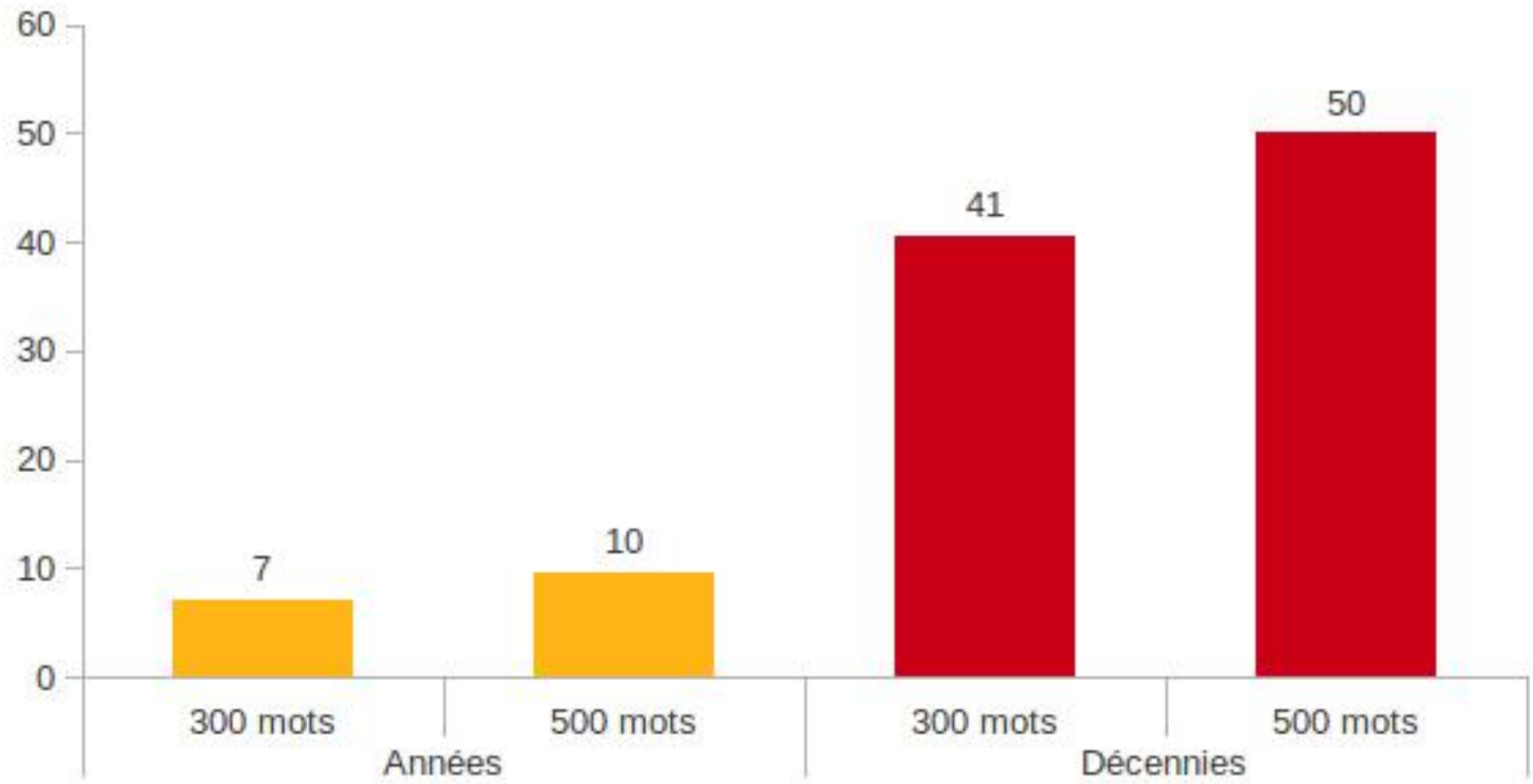
Résultats à DEFT 2011



Derniers résultats



Derniers résultats en termes de décennies et d'années correctes



Plan

Perspectives

Conclusion

- Meilleurs résultats avec SVM fondés sur *n*-grams de caractères + indices chronologiques
- Environ 10% d'années correctes, et 50% de décennies correctes (portions de 500 mots)

Perspectives

- Comparaison des indices fournis par chaque méthode
  - ex : date de naissances vs. Google n-grams

  $\rightarrow$ Étendre aux *n*-grams
- Correction orthographique
- Utiliser d'autres corpus

Évaluation du coût

- Nb de feutres : 5
- Nb de cafés : beaucoup trop
- Nb modifications wiki : près de 300
- Taux de refus essuyés lors de tentatives d'obtention d'informations privilégées auprès de Cyril : 100%

Évaluation du gain

- On s'est bien amusé(e)s
  - Nb de fous rires : supérieur à 20
- Publications :
  - 2 articles : DEFT et SPIRE 2011
  - 1 chapitre de livre : à venir