Construction et validation d'une base de données orthographiques

Jean-Luc MANGUIN
GREYC, CNRS UMR 6072, Caen

Collecter et produire des données pour la recherche en SHS
Fréjus, 15-18 novembre 2016

MATE-SHS

Introduction

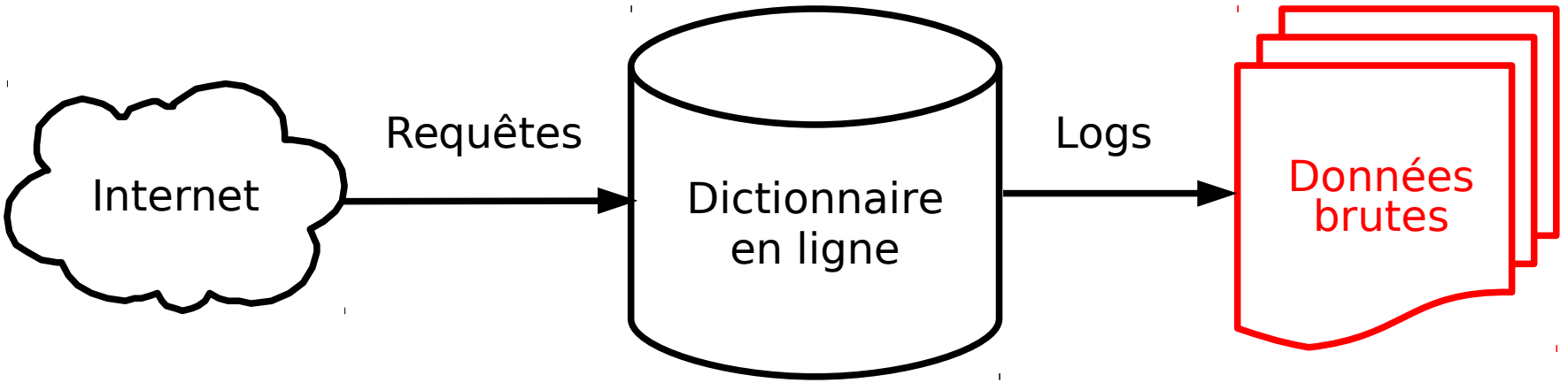
- Une base pour l'orthographe lexicale

Pour des besoins en psycholinguistique

- Un recueil de données original

Mais qui pose des problèmes de validation

Recueil des données



- Nécessite un accès administrateur

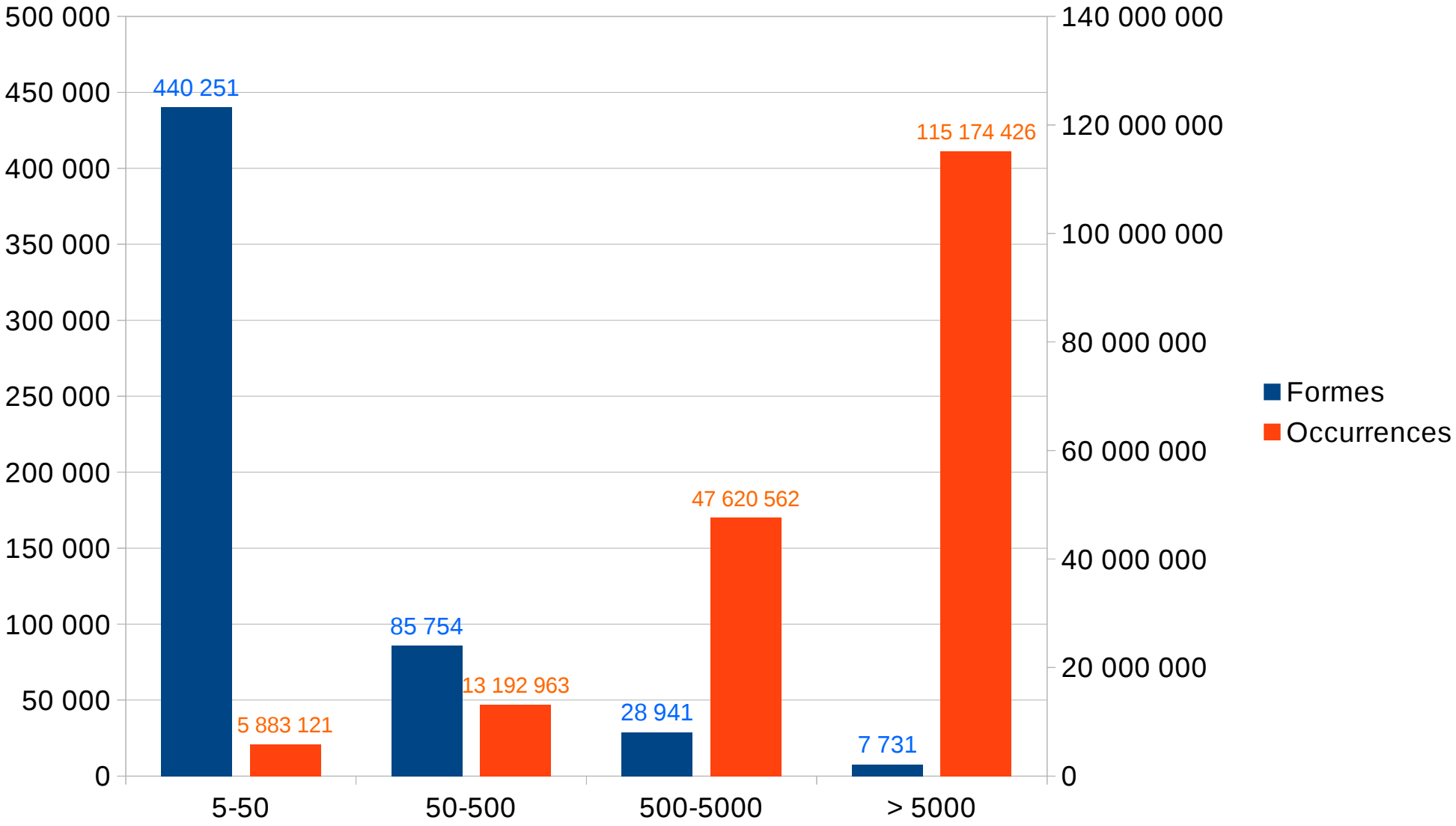
Et un traitement ultérieur

Un corpus de requêtes

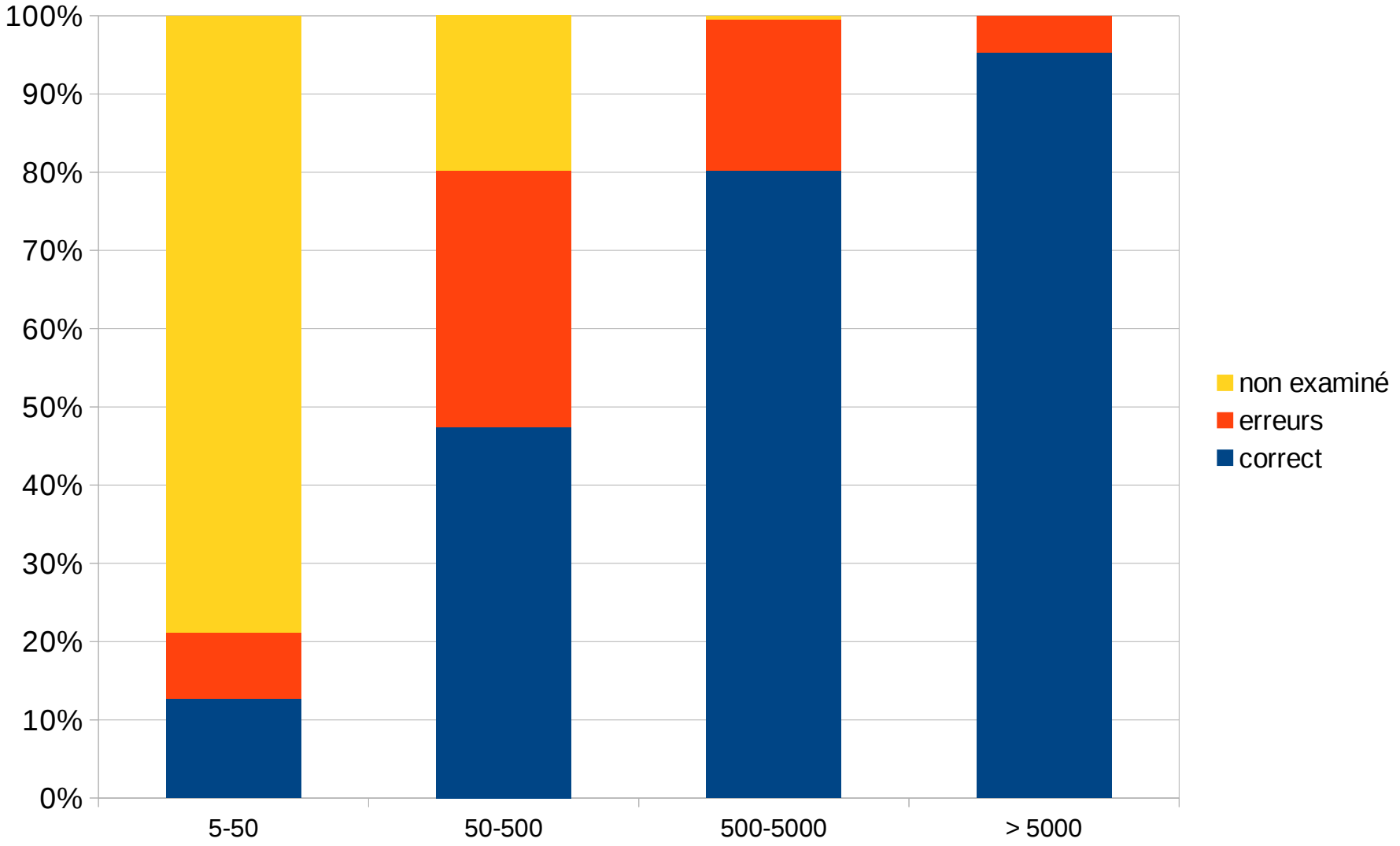
- Requêtes reçues par le dictionnaire des synonymes du CRISCO pendant 9 ans
- 188 millions d'occurrences – 3,6 millions de formes

  - MAIS (pour la base actuelle) :
  - On garde si f ≥ 5
  - 182 millions d'occurrences - 562 000 formes

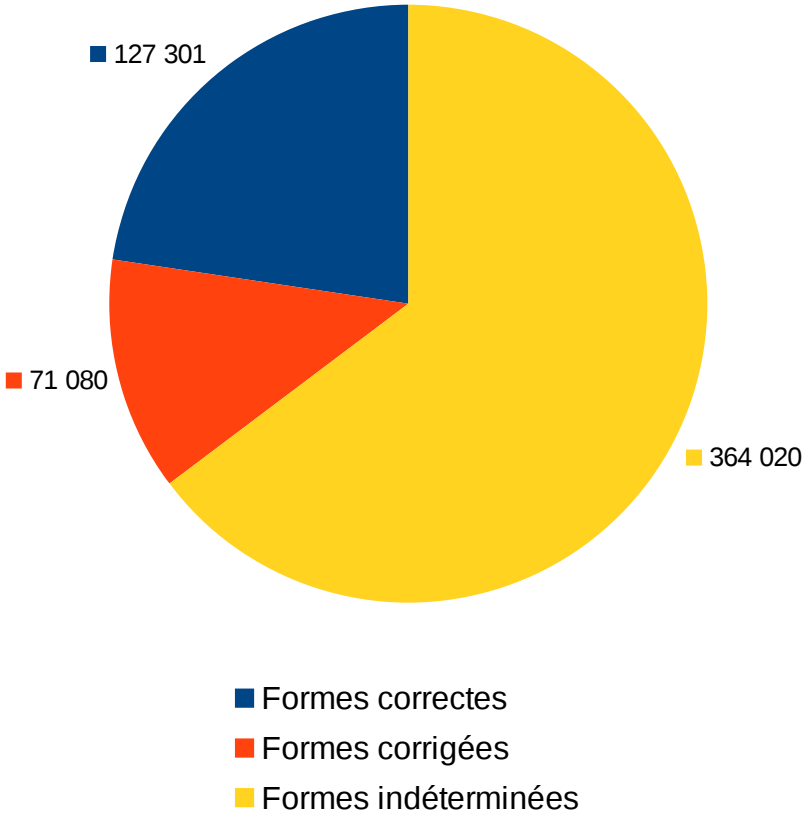
Répartition (formes et occurrences)

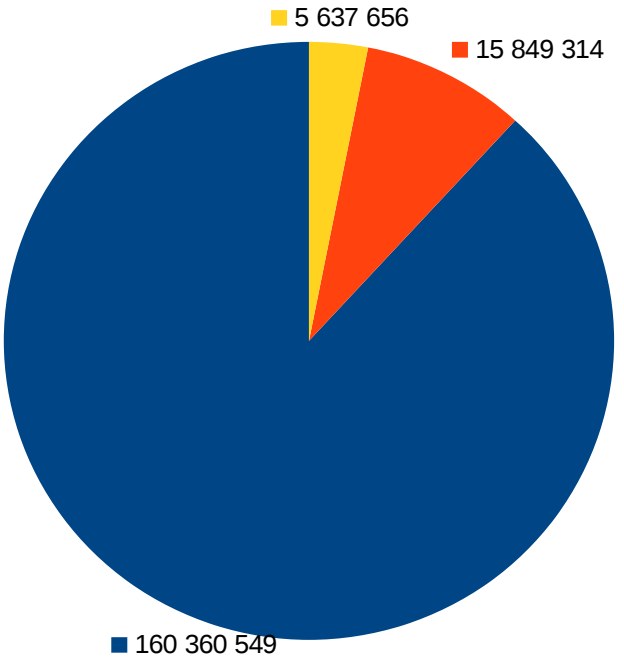


Répartition des erreurs



Autre synthèse





Validation de la base

- Pourquoi valider ?

  Notre méthode de recueil est inhabituelle

- Comment valider ?

  Par comparaison (quantitative, si possible) avec un corpus textuel, et avec une dictée.

MATE-SHS

Un corpus de textes

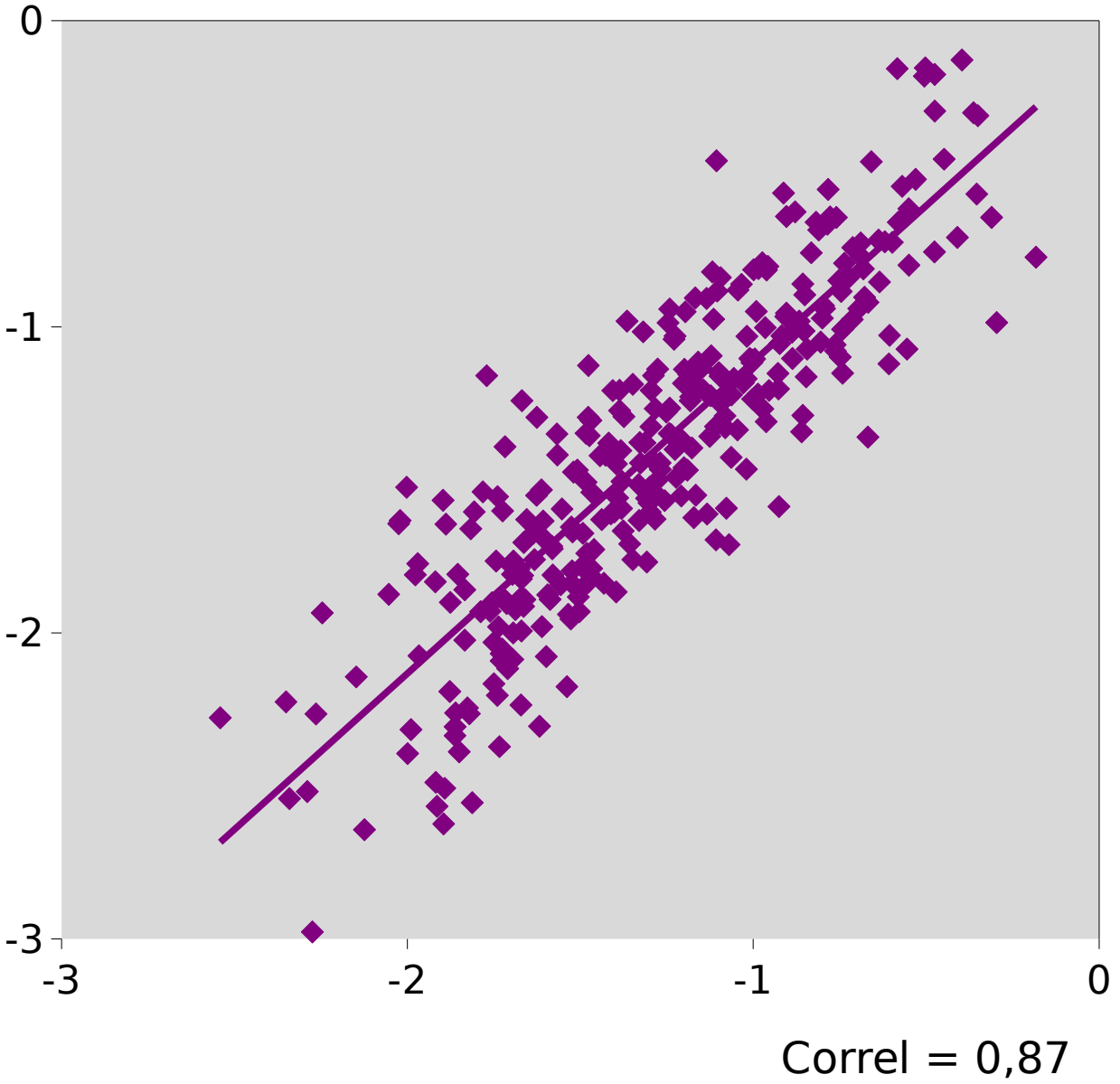
- Corpus « tapé » spontané
- Pas de spécialisation

- Forums divers « forum.*.fr »
- Interrogation manuelle par Google

Méthode de comparaison

- 350 formes contenant « onn[voyelle] »
- Relevé de la proportion de « on[voyelle] » pour chacune

La comparaison !



Un corpus de mots manuscrits

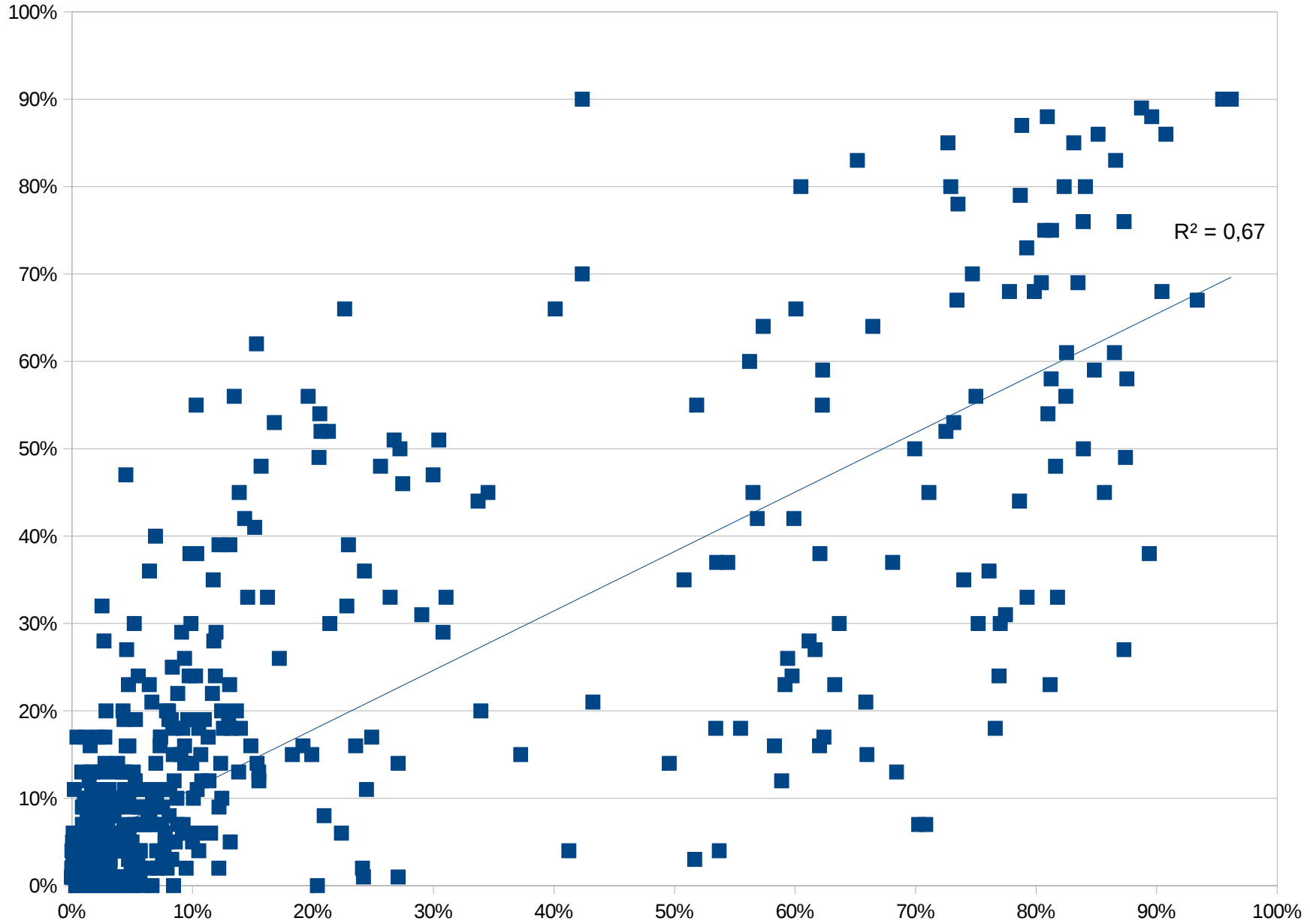
- 100 formes choisies
- Dictées à 100 sujets
- Soit 10 000 mots

Comparaison avec une dictée

- Proportion des formes (ex. pour la dictée)

commettre	8579	73,44%	67	67%
commetre	1534	13,13%	5	5%
comettre	1528	13,08%	23	23%
commaître	25	0,21%	2	2%
comètre	15	0,13%	2	2%
comaittre	1	0,01%	1	1%

Graphiquement



Correl = 0,82

Autre méthode

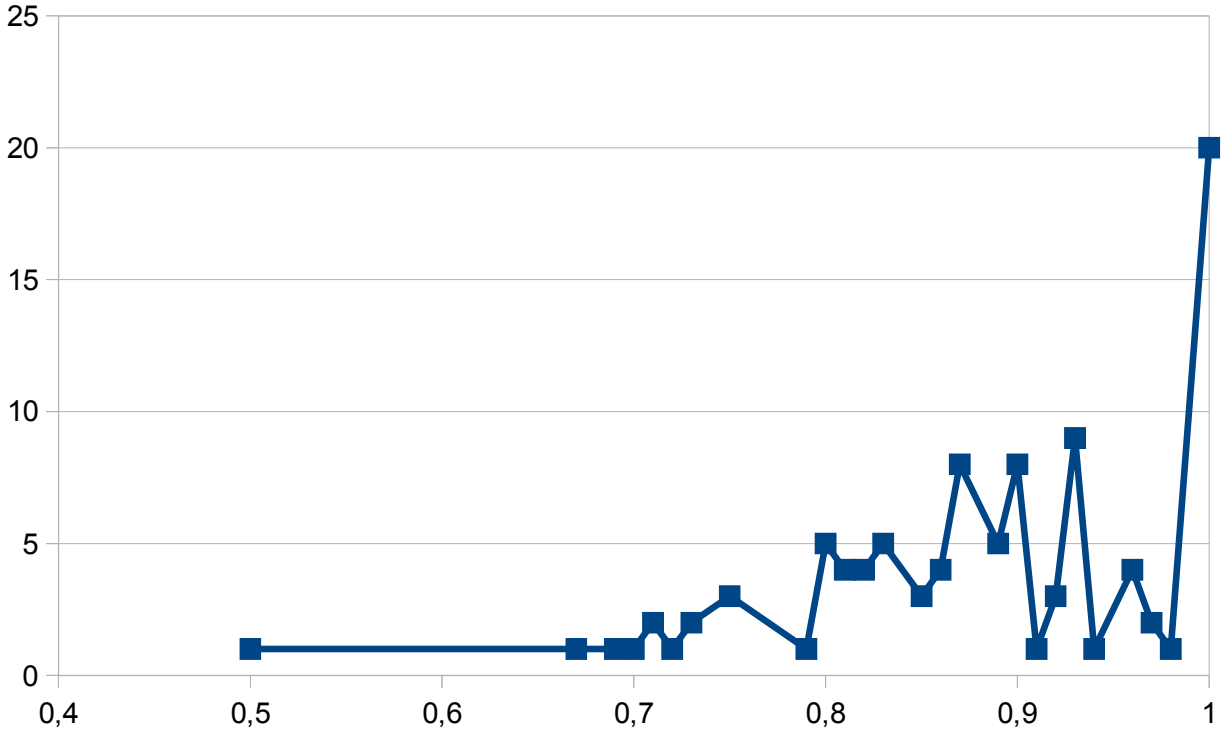
- Comparaison des rangs

		Rang 1		Rang 2
commettre	8579	1	67	1
commetre	1534	2	5	3
comettre	1528	3	23	2
commaître	25	4	2	4
comètre	15	5	2	5
comaittre	1	6	1	6

1 inversion sur 15 paires => accord à 93 %

Comparaison par rangs

- Sur l'ensemble des formes : accord à 88 %



Explication des différences

- Niveau des scripteurs

  Le DES est utilisé par des professionnels de la plume

- Accès à l'interface par copier-coller

  Depuis un logiciel qui possède déjà un correcteur

Avantages du recueil choisi

- Automatisation

  Pas de passation d'expérience

- Masse de données

  Observation de phénomènes plus rares

- Corpus « intensif »

  Gain en volume et en qualité

Inconvénients

- Techniques

  Accès au serveur Web et traitements

- Pas d'information contextuelle

  Recours nécessaire à des sources parallèles

Conclusion

- Adresse du site :

https://ortholexies.greyc.fr/

Merci de votre attention !

Avez-vous des questions ?