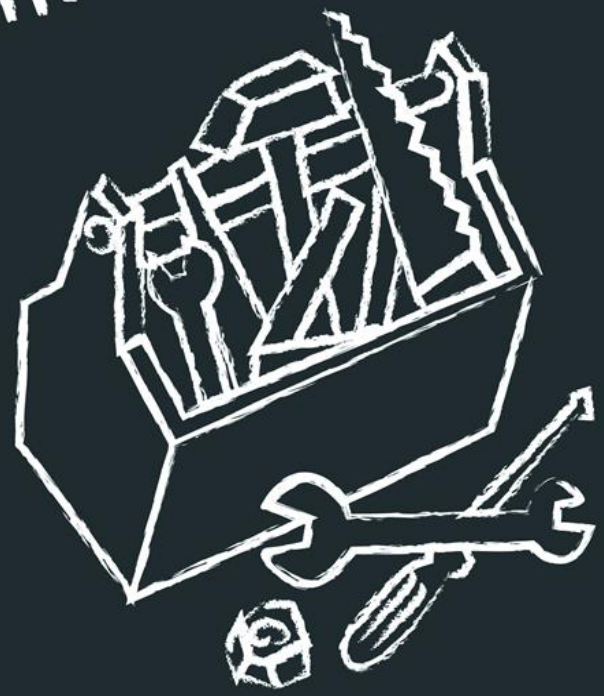
mshe
MAISON DES SCIENCES DE L'HOMME ET DE L'ENVIRONNEMENT
CLAUDE NICOLAS LEDOUX—USR 3124

LES TRUCS ET ASTUCES DE LA PLATE-FORME TECHNOLOGIQUE



De la bonne utilisation de Google Ngram Viewers

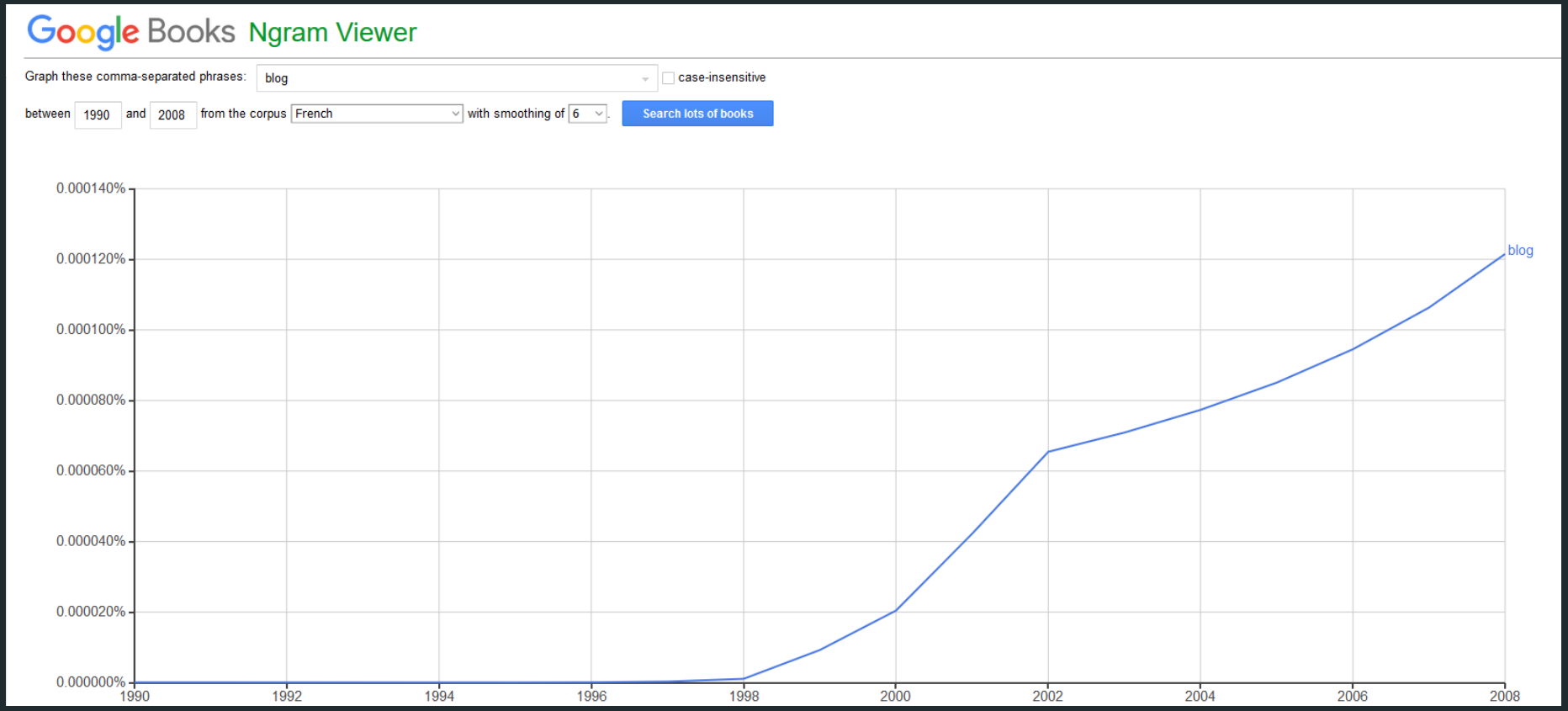
Nicolas Gutehrlé
nicolas.gutehrle@univ-fcomte.fr

PROGRAMME SUR
HTTPS://MSHE.UNIV-FCOMTE.FR

Introduction

- Application lancée le 16 décembre 2010.
- Repose sur le projet Google Books.
  - Environ 8 millions de livres numérisés.
  - 4% des livres publiés dans le monde.
  - Près de 500 milliards de mots .
- Outil de visualisation des fréquences d'utilisation d'un ou de plusieurs termes.
  - Voir l'évolution de leur utilisation dans le temps.
  - Voir le moment de leur apparition dans le corpus.

Apparition du mot « blog » en français



La culturonomique

- **Google Ngram Viewers** considéré comme à l'origine de **la culturonomique** (***culturomics*).**
  - Association de **culture** et **genomics.**
  - Nouveau champ d'application de la **lexicométrie.**
  - Vise à faire apparaître des phénomènes linguistiques et culturels par l'analyse de quantités gigantesques.
- Google Ngram n'est pas le premier outil à proposer des analyses statistiques sur la langue.
  - Outils existants déjà proposés dans le cadre de la **lexicométrie** et de la **textométrie.**
- Cependant, premier outil à appliquer ces techniques sur des corpus aussi importants.

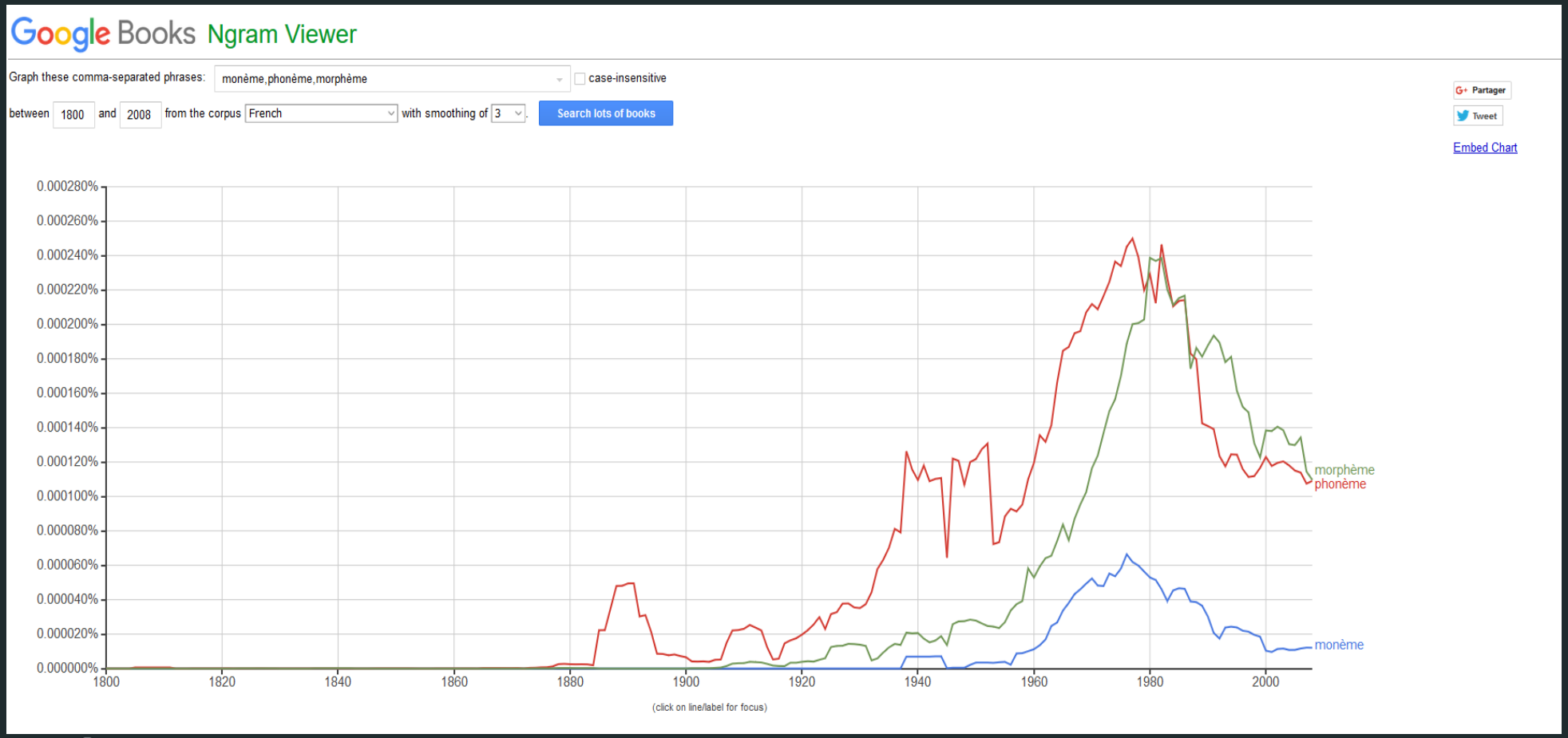
N-gram

- L'unité recherchée dans le corpus par Google Ngram Viewers est appelée **gramme (*gram*).**
  - Suite de caractères séparés par des espaces (« roi », « aujourd'hui », « porte-à-porte »).
- Un **n-gramme** est une suite de gramme de taille **n:**
  - **Unigramme :** éducation, démocratie.
  - **Bigramme :** éducation sentimentale, démocratie participative.
  - **Trigramme :** une nuit étoilée, hausse des prix.
  - **Tétragramme :** carnet de santé publique.
  - **Pentagramme :** principes fondamentaux de la philosophie.
- Google Ngram Viewers accepte au maximum les pentagrammes.

Une utilisation très simple

- Une barre de recherche dans laquelle on peut entrer les mots recherchés.
  - Par défaut, la recherche est sensible à la casse.
- Possibilité de choisir les dates de début et de fin du corpus.
  - De 1500 à 2012.
- Possibilité de choisir la langue du corpus (21 possibilités).
- Options pour choisir le degré de lissage de la courbe.
- Faire un clique droit sur la courbe affiche toutes les formes possibles du mot recherchés, avec leur courbes associées.
- On peut exporter les résultats :
  - Au format CSV.
  - Sur Twitter, Google+ et dans des pages web.

Une utilisation très simple



Des recherches plus sophistiquées

- Entrer un « * » à la place d'un mot fera apparaître les 10 substitutions les plus fréquentes.
  - Exemple : « université de * » au lieu de « université de Paris ».
- Cocher la case « case-insensitive » pour effectuer une recherche non-sensible à la casse.
- Pour rechercher un mot ainsi que toutes ses flexions, on peut ajouter **_INF** après ce mot.
  - Exemple : « livre_INF » au lieu de « livre ».
- Recherche à partir des **parties du discours :**
  - Chercher l'utilisation d'un mot pouvant avoir plusieurs catégories grammaticales.

    Exemple : (**livre**, verbe) et (**livre**, nom).

    livre_VERB, livre_NOUN
  - Chercher un motif.

    Exemple : lire _DET_ livre (lire un livre, lire le livre).

    Recherche limitées au trigrammes maximum.

Des recherches plus sophistiquées

❑ Parties du discours:

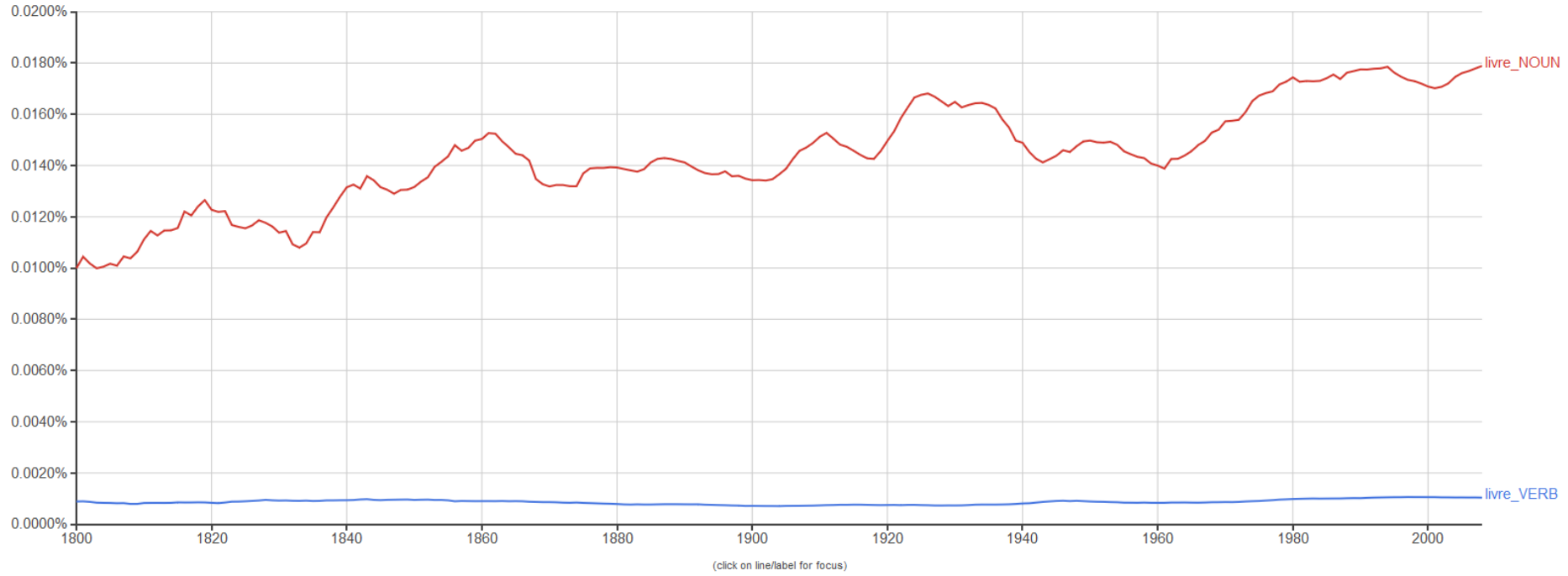
NOUN	Nom	Ces tags peuvent être utilisés seuls (_PRON_) ou en association avec un mot (elle_PRON).
VERB	Verbe	
ADJ	Adjectif	
ADV	Adverbe	
PRON	Pronom	
DET	Déterminant ou article	
ADP	Préposition ou postposition	
NUM	Numéral	
CONJ	Conjonction	
PRT	Particule	
ROOT	Racine de l'arbre d'analyse	Ces tags doivent être utilisés seuls (_START_).
START	Début de la phrase	
END	Fin de la phrase	

Des recherches plus sophistiquées

Google Books Ngram Viewer

Graph these comma-separated phrases: livre_VERB,livre_NOUN case-insensitive

between 1800 and 2008 from the corpus French with smoothing of 3 . Search lots of books



Des recherches plus sophistiquées

- Il est possible de combiner les différents tags:
  - Chercher les déterminants les plus fréquents allant avec le verbe « lire » .

    Exemple : **lire *_DET livre**
  - Chercher toutes les formes possibles d'un nom contenu dans un groupe nominal.

    Exemple : **livre_INF _NOUN_**

- On peut chercher les groupes dans lesquels un mot vient en modifier un autre:
  - On utilise l'opérateur **=>.**
  - Exemple : politiquement correct.

    politiquement => correct
  - On peut utiliser un * pour trouver tous les mots fréquemment associés à « politiquement ».

    politiquement => *

Des recherches plus sophistiquées

- Opérateurs pour la composition des **n-grammes** : + - / * :
  - **+** Additionne les fréquences d'apparition des n-grammes.

    Exemple : (immigration + migration) au lieu de « migration, immigration
  - **-** Soustrait les fréquences d'apparition des différents n-grammes.

    Exemple : (immigration – migration)
  - **/** Divise la fréquence d'apparition d'un n-gramme par celle d'un autre n-gramme.

    Exemple : (immigration / migration)
  - ***** Multiplie le n-gramme un nombre de fois donné. A utiliser pour visualiser des n-grammes dont les fréquences d'apparition sont très variables.

    Exemple : « (theremin*1000), violin » contre « theremin, violin »
  - **:** Compare en une seule requête les fréquences d'apparition des n-grammes dans différents corpus.

    Exemple : « (chat:fre_2012),(cat:eng_2012) »

Inconvénients

- Google Ngram Viewers ne permet que d'accéder aux résultats des calculs.
  - Impossible de contextualiser les résultats.
  - Pose d'importants soucis méthodologiques quant à l'interprétation des résultats.
  - Oblige l'analyste à faire appel à sa culture personnelle pour comprendre les résultats.
- Interaction avec Google Ngram n'est possible qu'en entrant les unités que l'on recherche.
  - Impossible de procéder à d'autres calculs que ceux de fréquences d'apparition.
- Des erreurs de numérisations mènent à des erreurs d'analyse.
  - Exemple : le mot « Internet » apparaissant pour la première fois en 1615 puis en 1990.
- La datation des ouvrages pose également problème.
  - Rééditions, collections, …

Conclusion

- Google Ngram Viewers permet des analyses linguistiques et statistiques sur un des plus grand corpus au monde.
- Outil intéressant pour initier un axe de recherche.
- Cependant, on ne peut qu'accéder aux résultats des calculs:
  - Impossible de revenir aux textes.
  - L'analyse des résultats doit être faite avec prudence.