Études quantitatives manuelles : problèmes d'échantillonnage dans les analyses métalexicographiques

UE TAL

Toulouse - 4/12/2023

Franck Sajous

CLLE (CNRS & Université de Toulouse 2)

CONTEXTE

Lexicographie, métalexicographie(s) et méthodologies

La lexicographie, c'est quoi ?

La lexicographie, c'est quoi ?

USITO UDS Le dictionnaire

lexicographie [lɛksikɔgʀafi] n. f.

Étude scientifique et analytique des faits de lexique d'une langue et de ses variétés dans le but de produire un dictionnaire.

Lexicographie française, québécoise.

Lexicographie bilingue.

TLFi

ÉTYMOLOGIE

1757; de lexico- et -graphie.

La lexicographie, c'est quoi ?

USITO UDS Le dictionnaire

lexicographie [lɛksikɔgʀafi] n. f.

Étude scientifique et analytique des faits de lexique d'une langue et de ses variétés dans le but de produire un dictionnaire.

Lexicographie française, québécoise.

Lexicographie bilingue.

TLFi

ÉTYMOLOGIE

1757; de lexico- et -graphie.

lexicographie [lɛksikɔgʀafi] **nom féminin**

ÉTYM. 1757 ◊ de *lexicographe*

■ LING. Travail et technique du lexicographe ; recensement et étude des mots et des expressions d'une langue déterminée, considérés dans leurs formes et leurs significations

La lexicographie, c'est quoi ?

USITO UDS Le dictionnaire

lexicographie [lɛksikɔgʀafi] n. f.

Étude scientifique et analytique des faits de lexique d'une langue et de ses variétés dans le but de produire un dictionnaire.

Lexicographie française, québécoise.

Lexicographie bilingue.

TLFi

ÉTYMOLOGIE

1757; de lexico- et -graphie.

~R

lexicographie [lɛksikɔgʀafi] **nom féminin**

ÉTYM. 1757 ◊ de *lexicographe*

■ LING. Travail et technique du lexicographe ; recensement et étude des mots et des expressions d'une langue déterminée, considérés dans leurs formes et leurs significations **(→ dictionnaire).** *Lexicographie et lexicologie*.*

La/les métalexicographie(s), c'est quoi ?

La/les métalexicographie(s), c'est quoi ?

Wiktionnaire
Le dictionnaire libre

Étymologie [modifier le wikicode]

(Date à préciser) Dérivé de *lexicographie*, avec le préfixe *méta-*.

Nom commun [modifier le wikicode]

métalexicographie \me.ta.le.ksi.ko.gʁa.fi\ *féminin*

Invariable
métalexicographie \me.ta.le.ksi.ko.gʁa.fi\

1. *(Linguistique)* Discipline qui étudie les méthodes et les principes guidant la création de dictionnaires.
   - *Nous ne retracerons pas ici les différents stades de l'évolution de la **métalexicographie** monolingue.* — (Witold Ucherek, *Les articles prépositionnels en lexicographie bilingue français-polonais*, 2019)

La/les métalexicographie(s), c'est quoi ?

Une variété d'activités/de sous-disciplines

metalexicography, lexicographic research, academic lexicography, dictionary research, theory of lexicography/theoretical lexicography, dictionary criticism/dictionary review, user research (user skills + user needs)

"it can be used by different authories to refer to potentially quite different things" (Hartmann, 2001, p. 28)

"The word *metalexicography* [...] is now frequently used to refer to the activities of anyone who writes about lexicography but does not write dictionaries" (Béjoint, 2000, p. 8)

La/les métalexicographie(s), c'est quoi ?

Une variété d'activités/de sous-disciplines

metalexicography, lexicographic research, academic lexicography, dictionary research, theory of lexicography/theoretical lexicography, dictionary criticism/dictionary review, user research (user skills + user needs)

"it can be used by different authories to refer to potentially quite different things" (Hartmann, 2001, p. 28)

"The word *metalexicography* [...] is now frequently used to refer to the activities of anyone who writes about lexicography but does not write dictionaries" (Béjoint, 2000, p. 8)

Définition maison (de la métalexicographie que je pratique)

Discipline qui consiste à étudier (décrire, analyser, évaluer, comparer) les dictionnaires et/ou leur processus de création...
indépendamment d'une visée particulière

(R)évolutions lexicographiques

Évolutions théoriques, technologiques et économiques

- révolution descriptive (Trap-Jensen, 2018)
- numérisation des dictionnaires papiers, rétroconversion vers BDD (Nagao et al., 1980; Berg et al., 1988)
- linguistique de corpus (Rundell & Stock, 1992)
  + automatisation par outils de TAL et d'analyse de données (Rundell & Kilgarriff, 2011)
- diversification des supports de publication, mise en ligne (Nesi, 2008)
- pour certains dictionnaires, arrêt de l'impression papier (Rundell, 2014)
- changement de modèle économique (Kilgarriff, 2005)
- apparition de dictionnaires « DIY » et d'agrégateurs (Gao, 2012)
- émergence de la lexicographie dite « collaborative » et des approches par *crowdsourcing* (Sajous & Josselin-Leray, 2022)

Quid de la métalexicographie ?

Une crise existentielle (de Schryver, 2022) ?

Ici, questionnement différent (car métalexicographie différente) :

- évolutions de la lexicographie $\Rightarrow$? changements/remise en cause de la méthode d'analyse métalexicographique
  ($\exists$? méthode d'analyse métalexicographique)
- pour un phénomène à étudier, quel type d'analyse privilégier ?
- quel impact du type de dictionnaire étudié sur les possibilités de mise en œuvre ?
  $\rightarrow$ analyse dictionnaire numérique $\approx$ analyse dictionnaire papier ?

réflexions d'ordre ~~*existentiel*~~ méthodologique : qu'est-ce qu'on fait et, surtout, ~~*pourquoi*~~ *comment* ?

Quid de la métalexicographie ?

Une crise existentielle (de Schryver, 2022) ?

Ici, questionnement différent (car métalexicographie différente) :

- évolutions de la lexicographie $\Rightarrow$? changements/remise en cause de la méthode d'analyse métalexicographique
  ($\exists$? méthode d'analyse métalexicographique)
- pour un phénomène à étudier, quel type d'analyse privilégier ?
- quel impact du type de dictionnaire étudié sur les possibilités de mise en œuvre ?
  $\rightarrow$ analyse dictionnaire numérique $\approx$ analyse dictionnaire papier ?

réflexions d'ordre ~~*existentiel*~~ méthodologique : qu'est-ce qu'on fait et, surtout, ~~*pourquoi*~~ *comment* ?

Suite de la présentation

- catégorisation des types d'analyse
- problèmes posés par chaque type d'analyse
- en particulier : problèmes posés par l'analyse d'échantillons

MÉTHODES D'ANALYSE

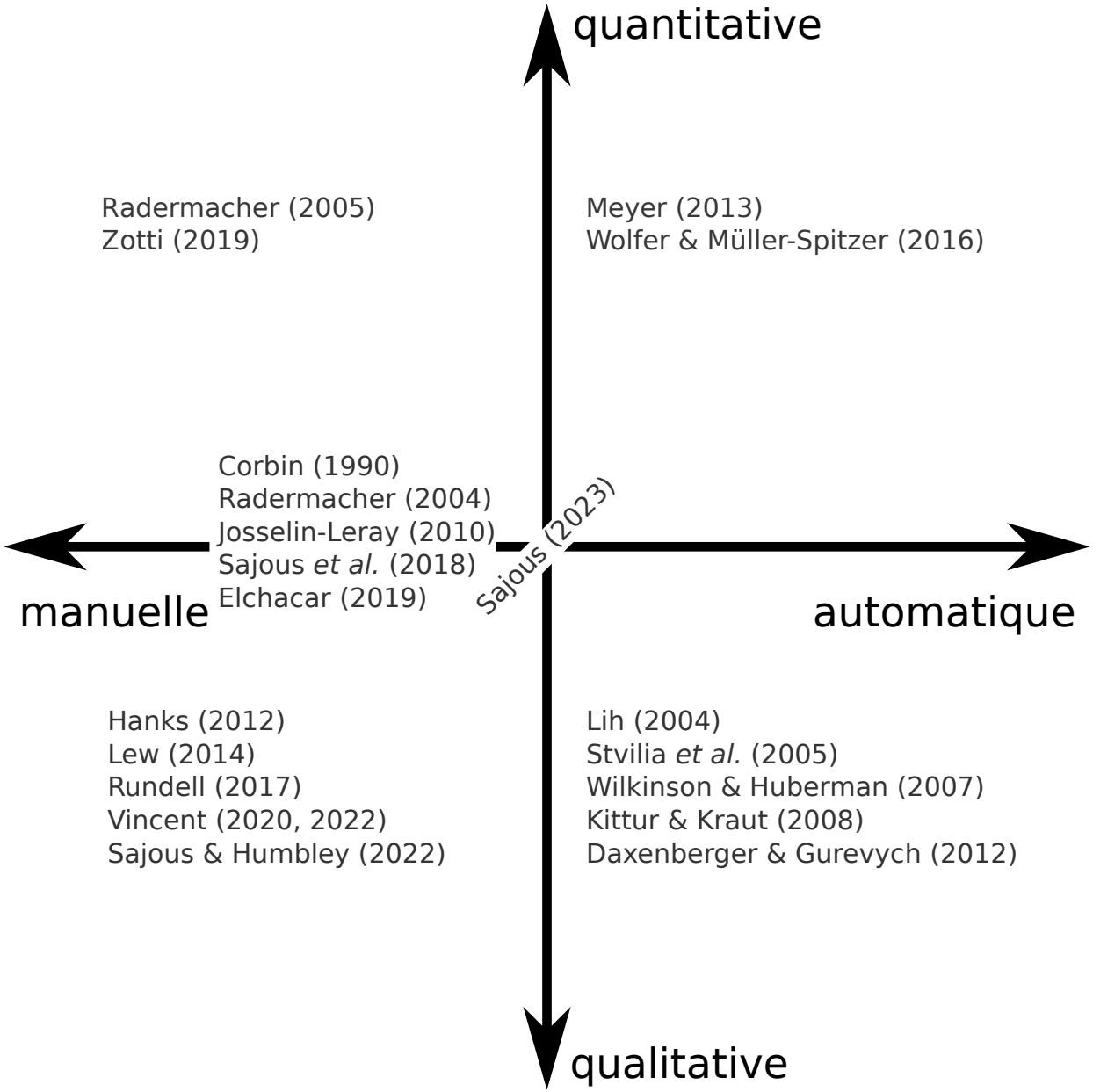
Éléments de classification

Méthodes d'analyse

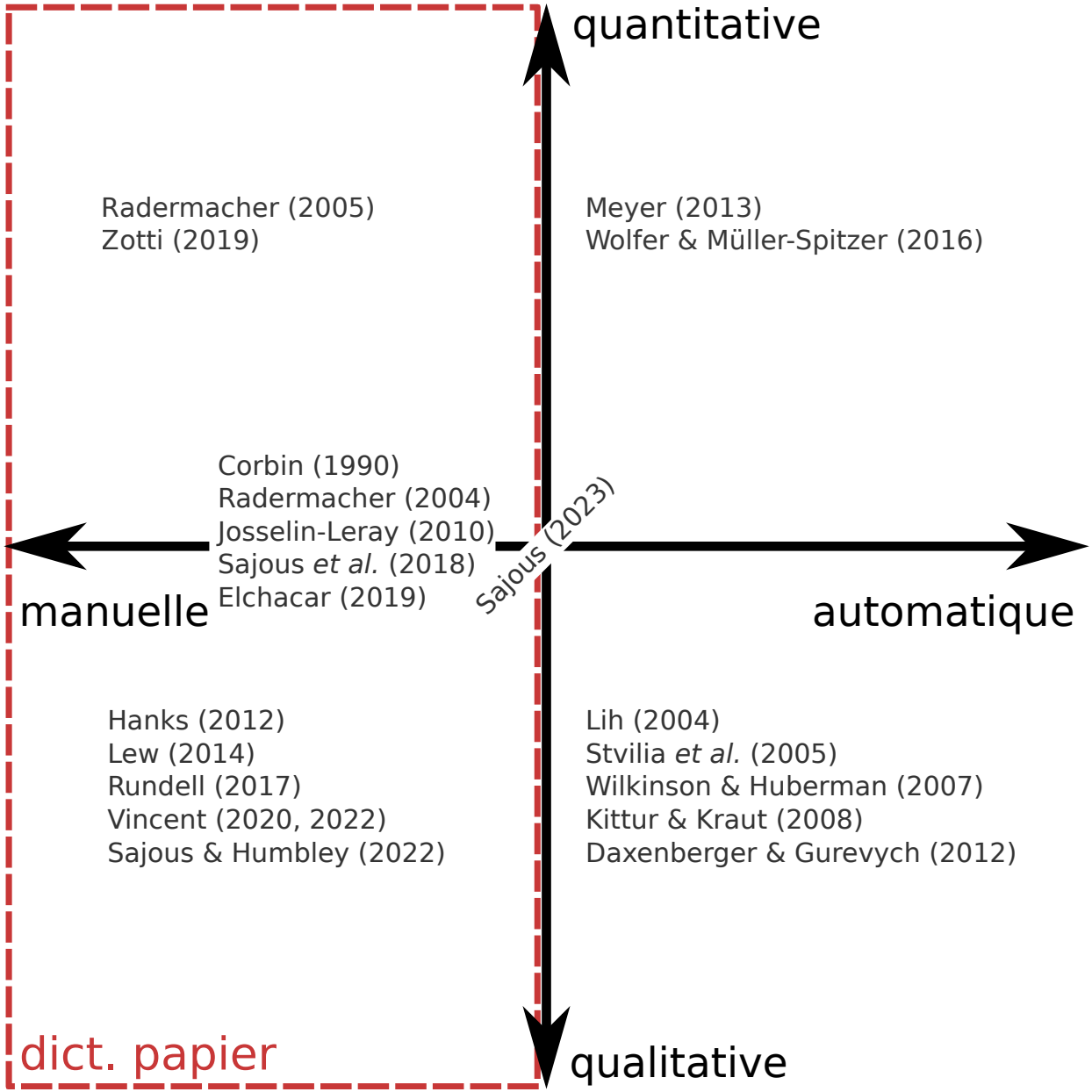
Éléments classificateurs retenus

1. Nature de l'analyse
   a. quantitative *vs.* qualitative
   b. synchronique *vs.* diachronique
2. Support du dictionnaire : papier *vs.* électronique
3. Profil des métalexicographes
4. Possibilité de mise en œuvre : manuelle *vs.* automatique

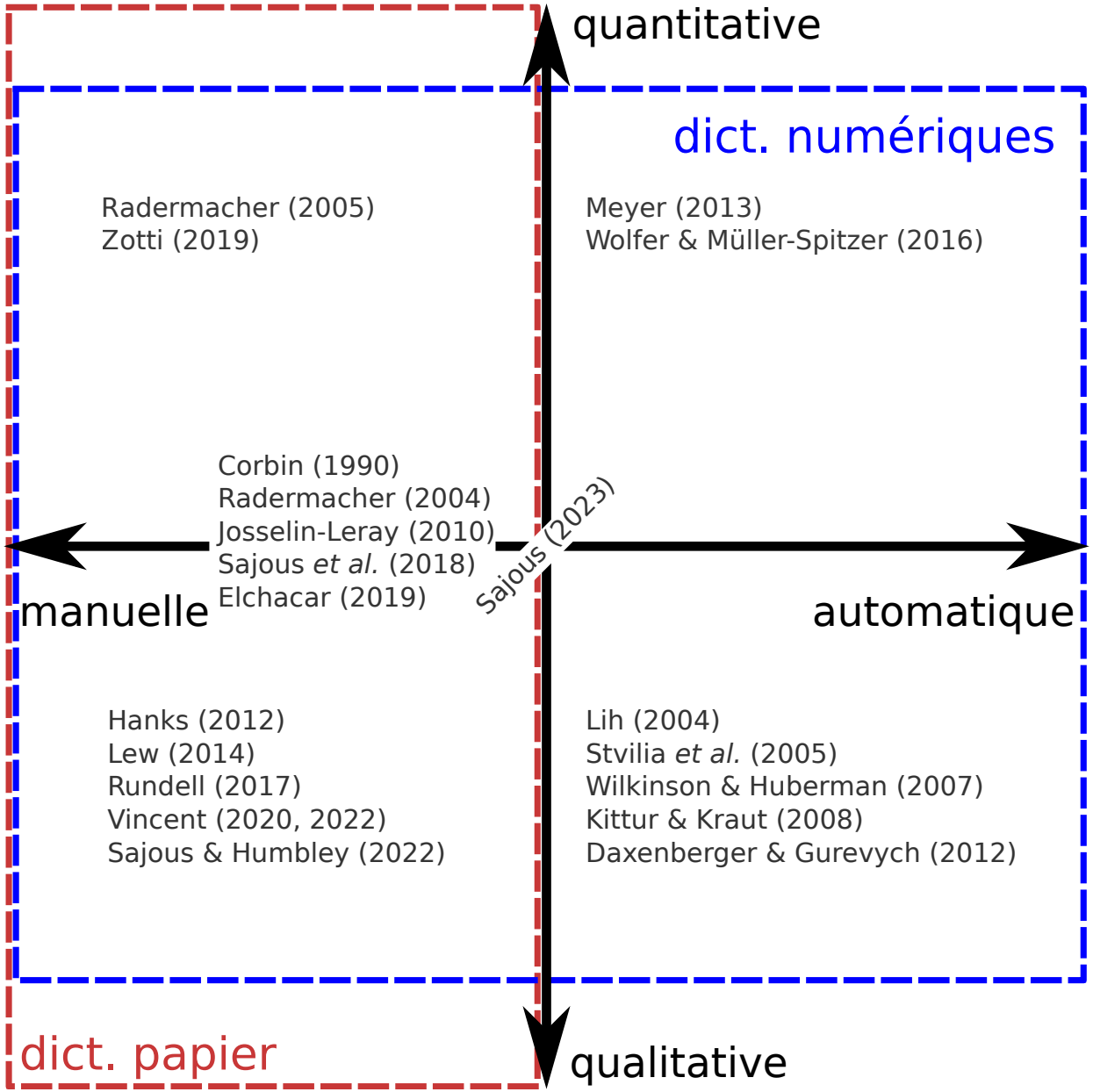
Méthodes d'analyse : axes (principaux) de classification



Méthodes d'analyse : axes (principaux) de classification



Méthodes d'analyse : axes (principaux) de classification



I - ANALYSES QUALITATIVES

1) manuelles

Analyses qualitatives manuelles

Analyses portant sur un nombre de cas (très) restreint

- Hanks (2012), Lew (2014), Rundell (2017)
  qualité des définitions dans *Wiktionary*
- Vincent (2022)
  traitement de l'entrée *woke* dans plusieurs dictionnaires
- Sajous & Humbley (2022)
  traitement d'entrées relatives aux mesures d'isolement sanitaire dans *Wiktionnaire* et *Wikipédia*

Qualitatif sans quantitatif : un problème ?

Non...

- études de cas illustrant des phénomènes observés, éventuellement de manière récurrente
- analyses fines possibles uniquement sur un nombre restreint d'exemples

mais :

- généralisation impossible à partir des observations

I - ANALYSES QUALITATIVES

2) automatiques

Automatiser des jugements qualitatifs ?

Démarche heuristique « faute de mieux »

- propriétés qualitatives difficiles à caractériser de manière computationnelle
- établissement *a priori* de critères souvent discutables

Exemple : étude de la qualité des articles de *Wikipédia*/*Wiktionary*

1. Relation avec le taux de révision (Stvilia et al., 2005; Wilkinson & Huberman, 2007; Kittur & Kraut, 2008; Daxenberger & Gurevych, 2012)
   Études fondées sur les labels de qualité attribués par les contributeurs

   → démarche problématique : évaluation interne, critères de qualité très discutables (Sajous, 2023)
2. Métriques dédiées (Stvilia et al., 2005)
   - longueur des articles, « fraîcheur » (*currency*) de l'information
   - degré de formalité de la langue
     → fréquence de POS *Wikipédia* ≈? *Columbia Encyclopedia*

Automatiser des jugements qualitatifs ?

Démarche heuristique « faute de mieux »

- propriétés qualitatives difficiles à caractériser de manière computationnelle
- établissement *a priori* de critères souvent discutables

Exemple : étude de la qualité des articles de *Wikipédia*/*Wiktionary*

1. Relation avec le taux de révision (Stvilia et al., 2005; Wilkinson & Huberman, 2007; Kittur & Kraut, 2008; Daxenberger & Gurevych, 2012)
   Études fondées sur les labels de qualité attribués par les contributeurs

   → démarche problématique : évaluation interne, critères de qualité très discutables (Sajous, 2023)
2. Métriques dédiées (Stvilia et al., 2005)
   - longueur des articles, « fraîcheur » (*currency*) de l'information
     - $\exists$ « bonne » longueur d'article, $\forall$ sujet ?
     - absence de mise à jour récente = prédicteur de mauvaise qualité ?
   - degré de formalité de la langue
     → fréquence de POS *Wikipédia* $\approx$? *Columbia Encyclopedia*
     - estimation discutable de la formalité de la langue
     - degré de formalité de la langue $\Rightarrow$ qualité des articles ?

Automatiser des jugements qualitatifs ?

Analyses qualitatives automatiques

= études quantitatives qui tentent d'appréhender des caractéristiques qualitatives. . .

menées par des informaticien.ne.s. . .

qui (souvent) connaissent peu leur objet d'étude

et qui (souvent) calculent tout ce qui est calculable

- caractérisation de certaines propriétés qualitatives pas toujours automatisable de manière satisfaisante
- indices/prédicteurs calculables pas toujours intéressants / pas toujours pertinents pour caractériser le phénomène étudié

II - ANALYSES QUANTITATIVES

1) automatiques

Analyses quantitatives automatiques

Principal problème : l'accès aux données

- *Wiktionary* : *dump* téléchargeable, sous licence libre (Meyer, 2013; Wolfer & Müller-Spitzer, 2016)
- dictionnaires commerciaux/institutionnels
  1. domaine public : OK si numérisé dans format exploitable (*e.g. XMLLittré*)
  2. contemporains : grande ou petite porte ?
     - *TLFi* : convention éventuellement possible avec l'ATILF
       ERSS / CLLE : depuis 2005
     - *Usito* : demander poliment, attendre (5 mois), re-demander poliment, être mis en attente, re-attendre (2 ans), puis finalement lire les conditions d'utilisation
- N.B. : accessible en ligne $\not\Rightarrow$ copyleft, copyright $\not\Rightarrow$ inexploitable

  $\rightarrow$ lire les mentions légales !

Autres problèmes ($\approx$ études qualitatives automatiques)

Profil des analystes, choix des observables, interprétations...

II - ANALYSES QUANTITATIVES

2) manuelles

II - ANALYSES QUANTITATIVES
2) manuelles

Mise en œuvre / portée du phénomène étudié

- ensemble restreint d'articles sélectionnés sur critères spécifiques
- globalité du dictionnaire

Analyse quantitative / constitution d'un sous-corpus

Requêtage via interface de recherche

- systèmes +/- sophistiqués
  - *Usito* : requêtes sur les vedettes, recherche d'appariement exact et suggestions par complétion automatique
  - *TLFi* : requêtes multicritères complexes
  - *DAF*, *PR* : "recherche avancée", plus de fonctionnalités qu'*Usito*, plus intuitive (mais moins puissante) que *TLFi*
- y recourir avec circonspection : pratique lexicographique non systématique/incohérente, codage instable, outils trop frustes
  - *e.g.* comptage du nb total d'emprunts au français dans l'*OED* en ligne (Coleman & Ogilvie, 2009) = combinaison de la 2^{e} et 3^{e} éditions
    → alternance de *French.*, *Fr.* et *F.* dans la rubrique étymologique.
    Requête "F." ramène aussi "f." (= *from*).
  - *PR* : marque LITTÉR. (= « termes des études littéraires » mais aussi les mots de « la langue écrite élégante »)
    Cohabitation avec LITTÉRATURE et LITT. (non documentées)

Analyse quantitative / constitution d'un sous-corpus

Recensions internes. . .

= listes de (sous-)vocabulaires spécifiques

- page d'accueil d'*Usito* : POS, particularismes (québécismes et francismes, réalités typiquement québécoises/françaises), anglicismes critiqués, etc.
- *Wiktionnaire* : différents types de lexiques (lexique de l'informatique, insultes, termes vieillis, etc.)

= danger

→ attention aux mécanismes (systématiques ou non) qui sous-tendent la macrostructure à partir de laquelle ces listes sont constituées !
E.g. : étude sur les appellations des identités de genre non traditionnelles dans les dictionnaires « professionnels et profanes » (Elchacar, 2019)

- comparaison chiffrée des nomenclatures de plusieurs ressources : *GDT* en tête
- (sous-)nomenclature du *Wiktionnaire* fondée sur son « vocabulaire LGBTIQ »
- 5 des 6 entrées considérées absentes du *Wiktionnaire* sont présentes lors de l'étude
  → conclusion fausse

Analyse quantitative / constitution d'un sous-corpus

Sélection par parcours exhaustif de la nomenclature

- Corbin (1990) + 45 étudiants (tous crédités)
  recherche de noms de végétaux en *-ier* et de leur base apparente
  dans 5 grands dict. généraux monolingues « de langue » et encyclopédiques
  → 249 noms (171 à 201 par dict.) dont les définitions sont ensuite analysées
  (entre 850 et 1000 définitions !)
- Sajous et al. (2018), d'après Martinez (2013)
  examen des 3 334 ajouts aux PR 2008-2017 pour sélectionner les entrées
  relevant du domaine de l'informatique
  → 120 articles à analyser (dont moins de la moitié marqués INFORM.)

Analyse quantitative / constitution d'un sous-corpus

Sélection par parcours exhaustif de la nomenclature

- Corbin (1990) + 45 étudiants (tous crédités)
  recherche de noms de végétaux en *-ier* et de leur base apparente
  dans 5 grands dict. généraux monolingues « de langue » et encyclopédiques
  → 249 noms (171 à 201 par dict.) dont les définitions sont ensuite analysées
  (entre 850 et 1000 définitions !)
- Sajous et al. (2018), d'après Martinez (2013)
  examen des 3 334 ajouts aux PR 2008-2017 pour sélectionner les entrées relevant du domaine de l'informatique
  → 120 articles à analyser (dont moins de la moitié marqués INFORM.)

Sélection sur critères externes

- *e.g.* nomenclatures de terminologies, recueils de variantes diatopiques, de faux-amis, glossaires d'argot, etc.
- exploitation de corpus, *e.g.* Josselin-Leray (2010)
  extraction de 110 termes FR + 110 termes EN
  à partir d'un corpus bilingue de vulgarisation en volcanologie
  → étude de l'inclusion et du traitement dans 2 dicts. FR, 2 EN et 2 bilingues

Analyse quantitative / ensemble du dictionnaire

échantillonnage nécessaire. . .

quantification d'un phénomène sur un nombre restreint d'articles,
puis généralisation

oui, mais comment ?

d'après Bukowska (2010) :

- beaucoup d'énergie consacrée à l'analyse des échantillons
- peu de réflexion accordée aux mécanismes de sélection des échantillons eux-mêmes

Analyse quantitative / ensemble du dictionnaire

échantillonnage nécessaire. . .

quantification d'un phénomène sur un nombre restreint d'articles,
puis généralisation

oui, mais comment ?

d'après Bukowska (2010) :

- beaucoup d'énergie consacrée à l'analyse des échantillons
- peu de réflexion accordée aux mécanismes de sélection des échantillons eux-mêmes

∃ méthodes +/- pertinentes

e.g. : image, échantillon = 25% pixels

Analyse quantitative / ensemble du dictionnaire

échantillonnage nécessaire. . .

quantification d'un phénomène sur un nombre restreint d'articles,
puis généralisation

oui, mais comment ?

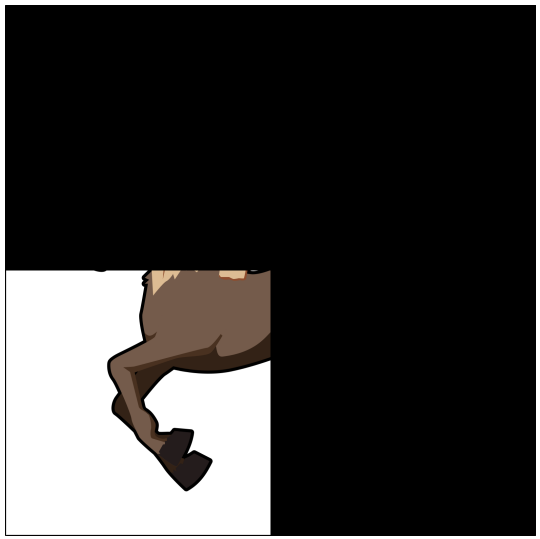
d'après Bukowska (2010) :

- beaucoup d'énergie consacrée à l'analyse des échantillons
- peu de réflexion accordée aux mécanismes de sélection des échantillons eux-mêmes

∃ méthodes +/- pertinentes

e.g. : image, échantillon = 25% pixels

- coin inférieur gauche



Analyse quantitative / ensemble du dictionnaire

échantillonnage nécessaire. . .

quantification d'un phénomène sur un nombre restreint d'articles,
puis généralisation

oui, mais comment ?

d'après Bukowska (2010) :

- beaucoup d'énergie consacrée à l'analyse des échantillons
- peu de réflexion accordée aux mécanismes de sélection des échantillons eux-mêmes

∃ méthodes +/- pertinentes

e.g. : image, échantillon = 25% pixels

- coin inférieur gauche
- coin inférieur droit

Analyse quantitative / ensemble du dictionnaire

échantillonnage nécessaire. . .

quantification d'un phénomène sur un nombre restreint d'articles,
puis généralisation

oui, mais comment ?

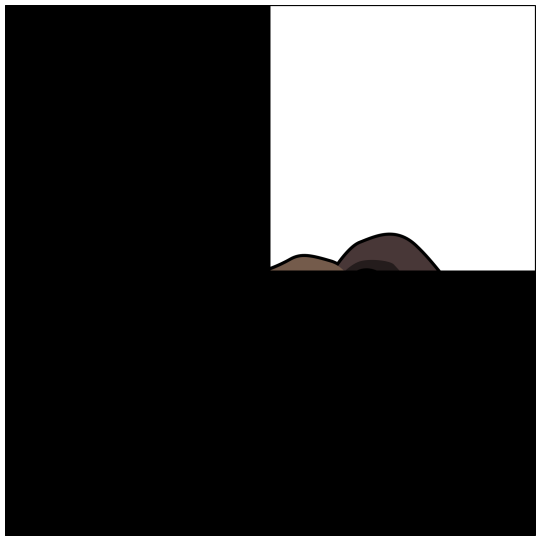
d'après Bukowska (2010) :

- beaucoup d'énergie consacrée à l'analyse des échantillons
- peu de réflexion accordée aux mécanismes de sélection des échantillons eux-mêmes

∃ méthodes +/- pertinentes

e.g. : image, échantillon = 25% pixels

- coin inférieur gauche
- coin inférieur droit
- coin supérieur droit



Analyse quantitative / ensemble du dictionnaire

échantillonnage nécessaire. . .

quantification d'un phénomène sur un nombre restreint d'articles,
puis généralisation

oui, mais comment ?

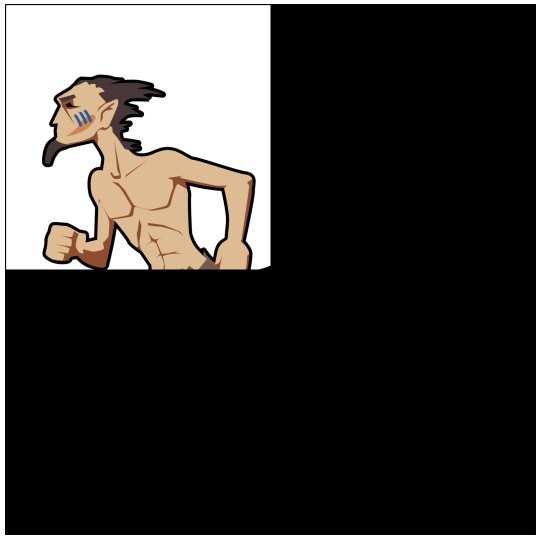
d'après Bukowska (2010) :

- beaucoup d'énergie consacrée à l'analyse des échantillons
- peu de réflexion accordée aux mécanismes de sélection des échantillons eux-mêmes

∃ méthodes +/- pertinentes

e.g. : image, échantillon = 25% pixels

- coin inférieur gauche
- coin inférieur droit
- coin supérieur droit
- coin supérieur gauche



Analyse quantitative / ensemble du dictionnaire

échantillonnage nécessaire. . .

quantification d'un phénomène sur un nombre restreint d'articles,
puis généralisation

oui, mais comment ?

d'après Bukowska (2010) :

- beaucoup d'énergie consacrée à l'analyse des échantillons
- peu de réflexion accordée aux mécanismes de sélection des échantillons eux-mêmes

∃ méthodes +/- pertinentes

e.g. : image, échantillon = 25% pixels

- coin inférieur gauche
- coin inférieur droit
- coin supérieur droit
- coin supérieur gauche
- tirage aléatoire

Analyse quantitative / ensemble du dictionnaire

échantillonnage nécessaire. . .

quantification d'un phénomène sur un nombre restreint d'articles,
puis généralisation

oui, mais comment ?

d'après Bukowska (2010) :

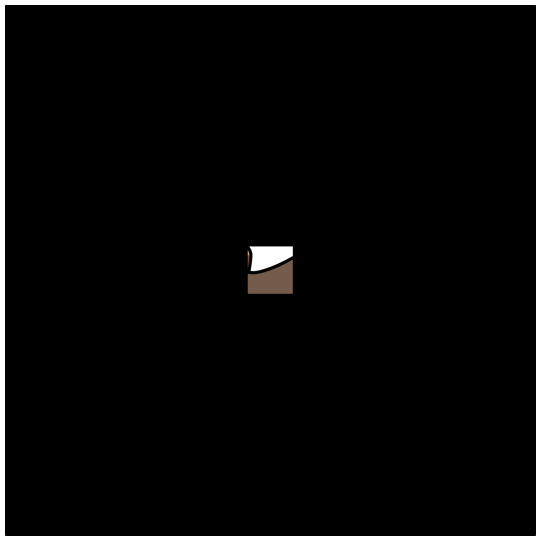
- beaucoup d'énergie consacrée à l'analyse des échantillons
- peu de réflexion accordée aux mécanismes de sélection des échantillons eux-mêmes

∃ méthodes +/- pertinentes

e.g. : image, échantillon = 25% pixels

- coin inférieur gauche
- coin inférieur droit
- coin supérieur droit
- coin supérieur gauche
- tirage aléatoire

En métalexicographie, échantillon ≈ 1% dict.



Analyse quantitative / ensemble du dictionnaire

échantillonnage nécessaire. . .

quantification d'un phénomène sur un nombre restreint d'articles,
puis généralisation

oui, mais comment ?

d'après Bukowska (2010) :

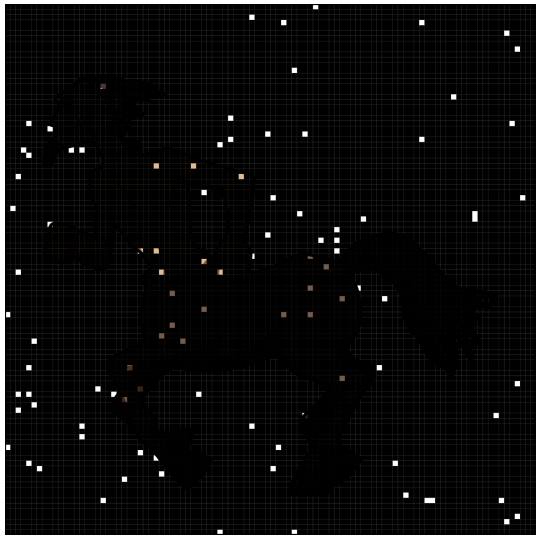
- beaucoup d'énergie consacrée à l'analyse des échantillons
- peu de réflexion accordée aux mécanismes de sélection des échantillons eux-mêmes

∃ méthodes +/- pertinentes

e.g. : image, échantillon = 25% pixels

- coin inférieur gauche
- coin inférieur droit
- coin supérieur droit
- coin supérieur gauche
- tirage aléatoire

En métalexicographie, échantillon ≈ 1% dict.



Tranche contiguë : biais importants

- propriétés inhérentes au lexique
- artefacts liés au processus rédactionnel
  - différences interpersonnelles de pratiques entre rédacteurs, changement de direction éditoriale, de type de corpus, etc.
  - *"alphabet fatigue"* (Osselton, 2007) : lettres du milieu de l'alphabet mieux adaptées à l'analyse car adoption d'un *modus operandi* régulier, entre rodage initial et accélération finale
    $\rightarrow$ présupposé discutable (de Schryver, 2005)

Méthodes d'échantillonnage

Tranche contiguë : biais importants

- propriétés inhérentes au lexique
- artefacts liés au processus rédactionnel
  - différences interpersonnelles de pratiques entre rédacteurs, changement de direction éditoriale, de type de corpus, etc.
  - *"alphabet fatigue"* (Osselton, 2007) : lettres du milieu de l'alphabet mieux adaptées à l'analyse car adoption d'un *modus operandi* régulier, entre rodage initial et accélération finale
    → présupposé discutable (de Schryver, 2005)

Échantillonnage probabiliste

- moins problématique car biais moins systématiques
- mais aucune garantie qu'un échantillon (même de taille raisonnable) soit représentatif

Méthodes d'échantillonnage

Tranche contiguë : biais importants

- propriétés inhérentes au lexique
- artefacts liés au processus rédactionnel
  - différences interpersonnelles de pratiques entre rédacteurs, changement de direction éditoriale, de type de corpus, etc.
  - *"alphabet fatigue"* (Osselton, 2007) : lettres du milieu de l'alphabet mieux adaptées à l'analyse car adoption d'un *modus operandi* régulier, entre rodage initial et accélération finale
    → présupposé discutable (de Schryver, 2005)

Échantillonnage probabiliste

- moins problématique car biais moins systématiques
- mais aucune garantie qu'un échantillon (même de taille raisonnable) soit représentatif

Et pourtant...

échantillonnage par tranche contiguë de loin le plus employé par les métalexicographes

Autres méthodes d'échantillonnage

Échantillonnage probabiliste stratifié

- découpage du dict. en zones non chevauchantes (les strates)

  e.g. les lettres initiales, les tomes, les parties rédigées par différents éditeurs

- échantillonnage aléatoire respectant certaines proportions :
  - celles du dictionnaire (*e.g.* POS, marquage, etc.)
  - calculées en corpus (*e.g.* POS, rangs de fréquence, etc.)

Autres méthodes d'échantillonnage

Échantillonnage probabiliste stratifié

- découpage du dict. en zones non chevauchantes (les strates)

  e.g. les lettres initiales, les tomes, les parties rédigées par différents éditeurs
- échantillonnage aléatoire respectant certaines proportions :
  - celles du dictionnaire (*e.g.* POS, marquage, etc.)
  - calculées en corpus (*e.g.* POS, rangs de fréquence, etc.)
- plus complexe, pas toujours possible, pas toujours pertinent :
  - proportions d'une caractéristique du dictionnaire souvent inconnues
  - n'améliore pas systématiquement l'échantillonnage probabiliste simple
  - aider/forcer le hasard ? construction artificielle/arbitraire d'un échantillon « trop » équilibré $\Rightarrow$? biais

Autres méthodes d'échantillonnage

Échantillonnage probabiliste stratifié

- découpage du dict. en zones non chevauchantes (les strates)

  e.g. les lettres initiales, les tomes, les parties rédigées par différents éditeurs
- échantillonnage aléatoire respectant certaines proportions :
  - celles du dictionnaire (*e.g.* POS, marquage, etc.)
  - calculées en corpus (*e.g.* POS, rangs de fréquence, etc.)
- plus complexe, pas toujours possible, pas toujours pertinent :
  - proportions d'une caractéristique du dictionnaire souvent inconnues
  - n'améliore pas systématiquement l'échantillonnage probabiliste simple
  - aider/forcer le hasard ? construction artificielle/arbitraire d'un échantillon « trop » équilibré $\Rightarrow$? biais

Échantillonnage systématique

sélection d'un observable tous les N $\rightarrow$ échantillonnage non aléatoire

la théorie des probabilités et les statistiques inférentielles ont peu à dire sur la confiance que l'on peut accorder à un échantillonnage non aléatoire

(Freeman, 1963) cité par Bukowska (2010)

À suivre. . .

Expériences d'échantillonnage

- génération automatique d'échantillons
- simulation de ce que les métalexicographes seraient susceptibles de faire manuellement
- comparaison des résultats obtenus avec les méthodes d'échantillonnage probabiliste *vs.* par zone contiguë
- questionnement sur la « fiabilité » des échantillons (et des moyens d'estimer cette fiabilité)

À suivre. . .

Expériences d'échantillonnage

- génération automatique d'échantillons
- simulation de ce que les métalexicographes seraient susceptibles de faire manuellement
- comparaison des résultats obtenus avec les méthodes d'échantillonnage probabiliste *vs.* par zone contiguë
- questionnement sur la « fiabilité » des échantillons (et des moyens d'estimer cette fiabilité)

Précisions

- les métalexicographes ne « connaissent » pas leur population
- les métalexicographes constituent et analysent UN échantillon (ou deux)

À suivre. . .

Expériences d'échantillonnage

- génération automatique d'échantillons
- simulation de ce que les métalexicographes seraient susceptibles de faire manuellement
- comparaison des résultats obtenus avec les méthodes d'échantillonnage probabiliste *vs.* par zone contiguë
- questionnement sur la « fiabilité » des échantillons (et des moyens d'estimer cette fiabilité)

Précisions

- les métalexicographes ne « connaissent » pas leur population
- les métalexicographes constituent et analysent UN échantillon (ou deux)

Disclaimer

- les expériences menées dans cette études ne sont pas celles que je préconise (ne faites pas ça chez vous !)
- je procède comme les métalexicographes sont susceptibles de le faire

III - EXPÉRIENCES D'ÉCHANTILLONNAGE

Type d'étude / dictionnaire analysé

1. **Étude synchronique : *Usito***
2. Étude diachronique : tomes du TLF

Estimation du marquage dans *Usito*

Usito

- dictionnaire général, « de langue », normatif, du français québécois
- réalisé par l'Université de Sherbrooke
- nativement numérique, en ligne
- gratuit depuis octobre 2019

Marques *fig.* et *fam.*

- proportion d'articles portant la marque *fig.* dans *Usito*
- proportion d'articles portant la marque *fam.* dans *Usito*

→ phénomènes choisis parmi beaucoup d'autres possibles, mais observables factuels (identifiables automatiquement), présence non anecdotique, répartition *a priori* sur l'ensemble du dictionnaire

Corpus et méthode

- restriction de l'étude aux noms → 31 310 articles concernés
- analyse automatique sur l'ensemble du dictionnaire → calcul % réel
- expériences automatiques d'échantillonnage → simulation de résultats obtenus par analyse manuelle

Échantillonnage : zone contiguë *vs.* tirage aléatoire

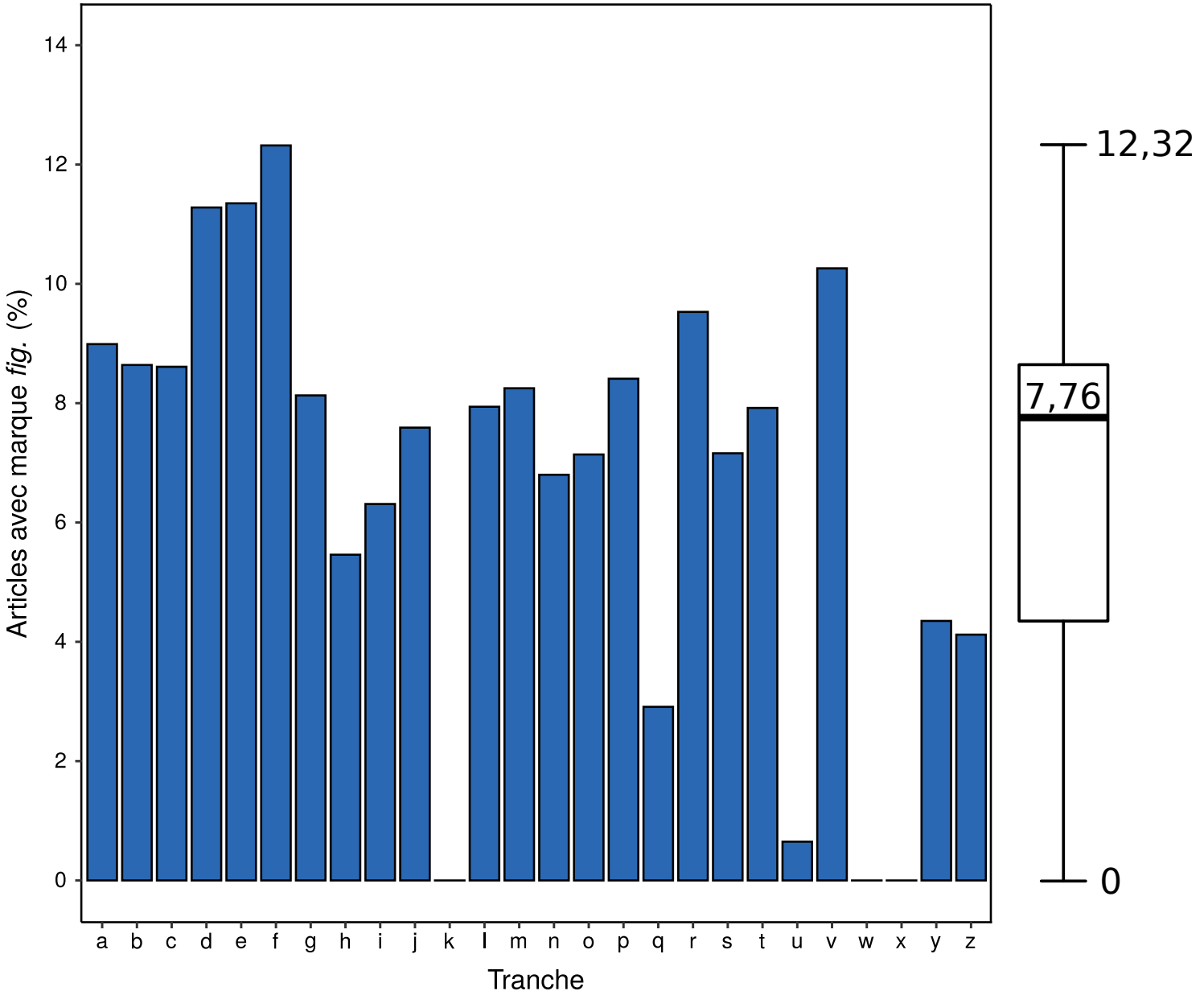
Tranche = vedettes commençant par la même lettre initiale

- sélection contiguë + tirage aléatoire d'échantillons de *n* articles dans une tranche donnée
- quelle tranche choisir ? Quelle incidence de ce choix ?

Tranche	Nb articles	% articles
a	3350	10,70
b	1909	6,10
c	3855	12,31
d	1826	5,83
e	1004	3,21
f	1234	3,94
g	1156	3,69
h	842	2,69
i	1061	3,39
j	290	0,93
k	165	0,53
l	945	3,02
m	2084	6,66

Tranche	Nb articles	% articles
n	647	2,07
o	700	2,24
p	3248	10,37
q	172	0,55
r	1669	5,33
s	2305	7,36
t	1667	5,32
u	154	0,49
v	799	2,55
w	61	0,19
x	24	0,08
y	46	0,15
z	97	0,31

Échantillonnage : zone contiguë *vs.* tirage aléatoire



Tranche	% *fig.*
a	8,99
b	8,64
c	8,61
d	11,28
e	11,35
f	12,32
g	8,13
h	5,46
i	6,31
j	7,59
k	0,00
l	7,94
m	8,25
n	6,80
o	7,14
p	8,41
q	2,91
r	9,53
s	7,16
t	7,92
u	0,65
v	10,26
w	0,00
x	0,00
y	4,35
z	4,12

Échantillonnage : zone contiguë *vs.* tirage aléatoire

Génération automatique des échantillons

Pour chaque tranche :

- génération de 100 échantillons de 500 (max.) articles contigus, articles de départ tirés aléatoirement
- génération de 100 échantillons de 500 (max.) articles, tous tirés aléatoirement

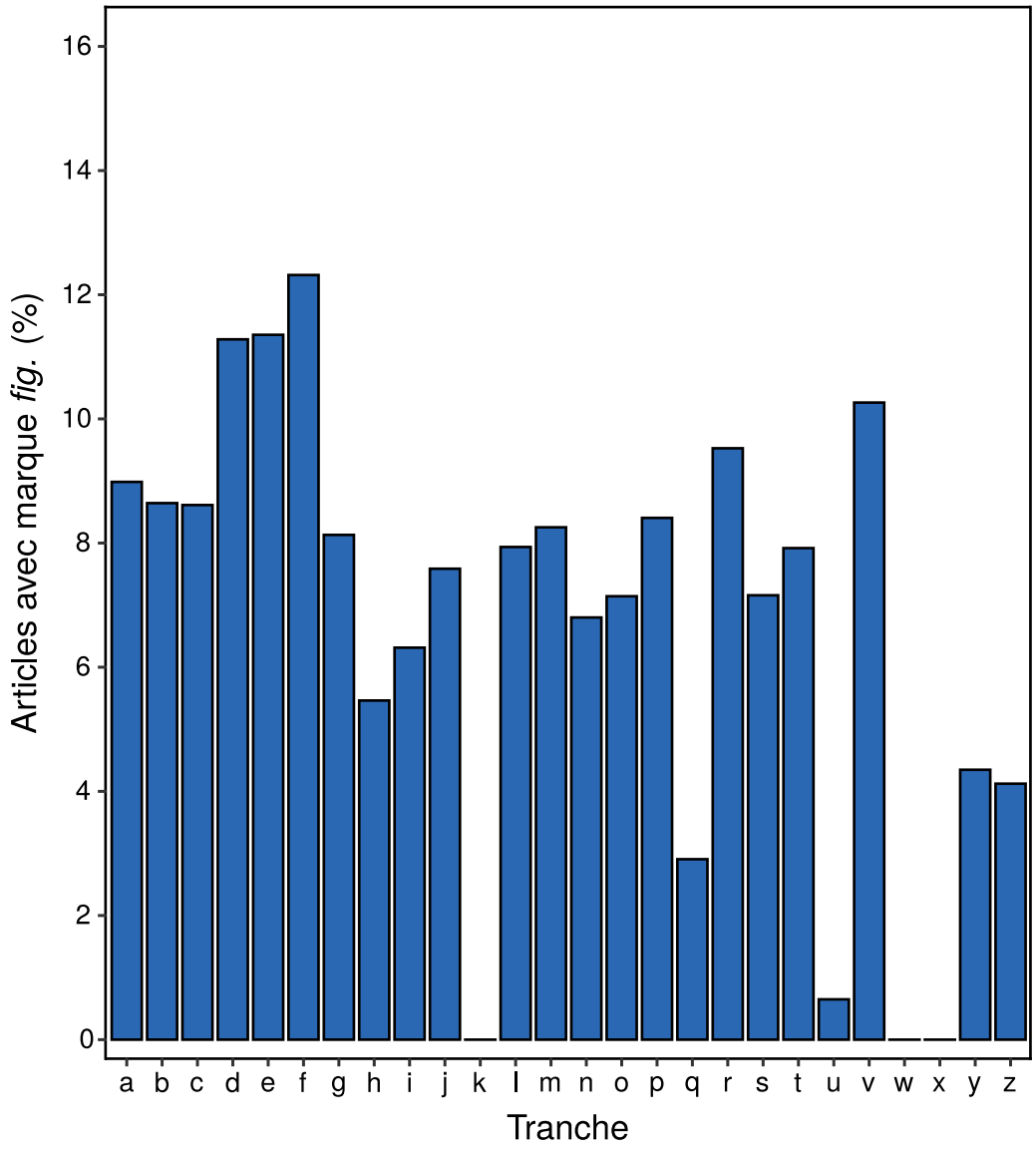
Pour chaque échantillon :

- calcul du % d'articles marqués

Observation des distributions des valeurs obtenues

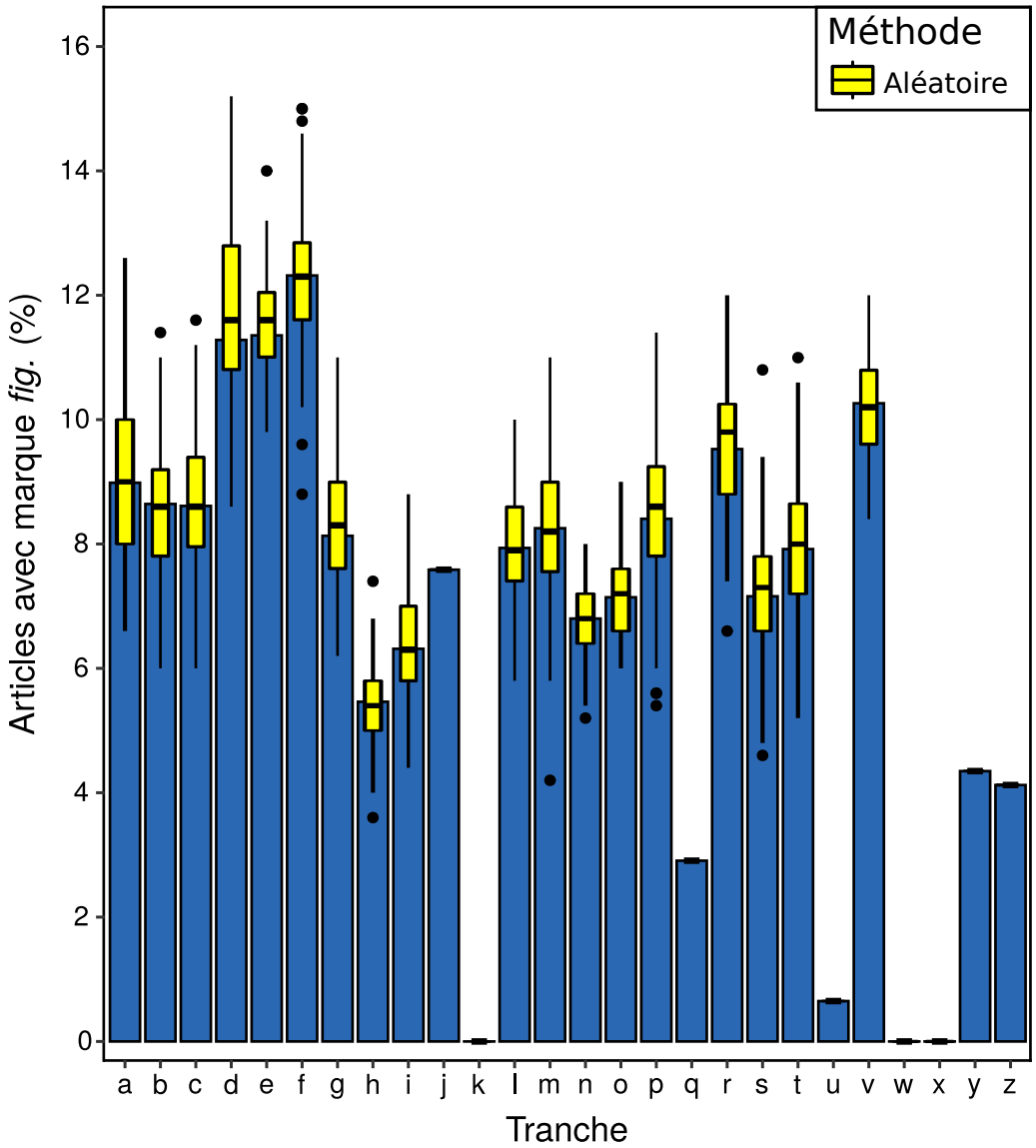
- quelle variabilité entre % échantillons ?
- quel écart entre % échantillons et % réel de la tranche ?
- quel écart entre % échantillons et % réel du dictionnaire ?

Échantillonnage : zone contiguë *vs.* tirage aléatoire



% marquage réel par tranche

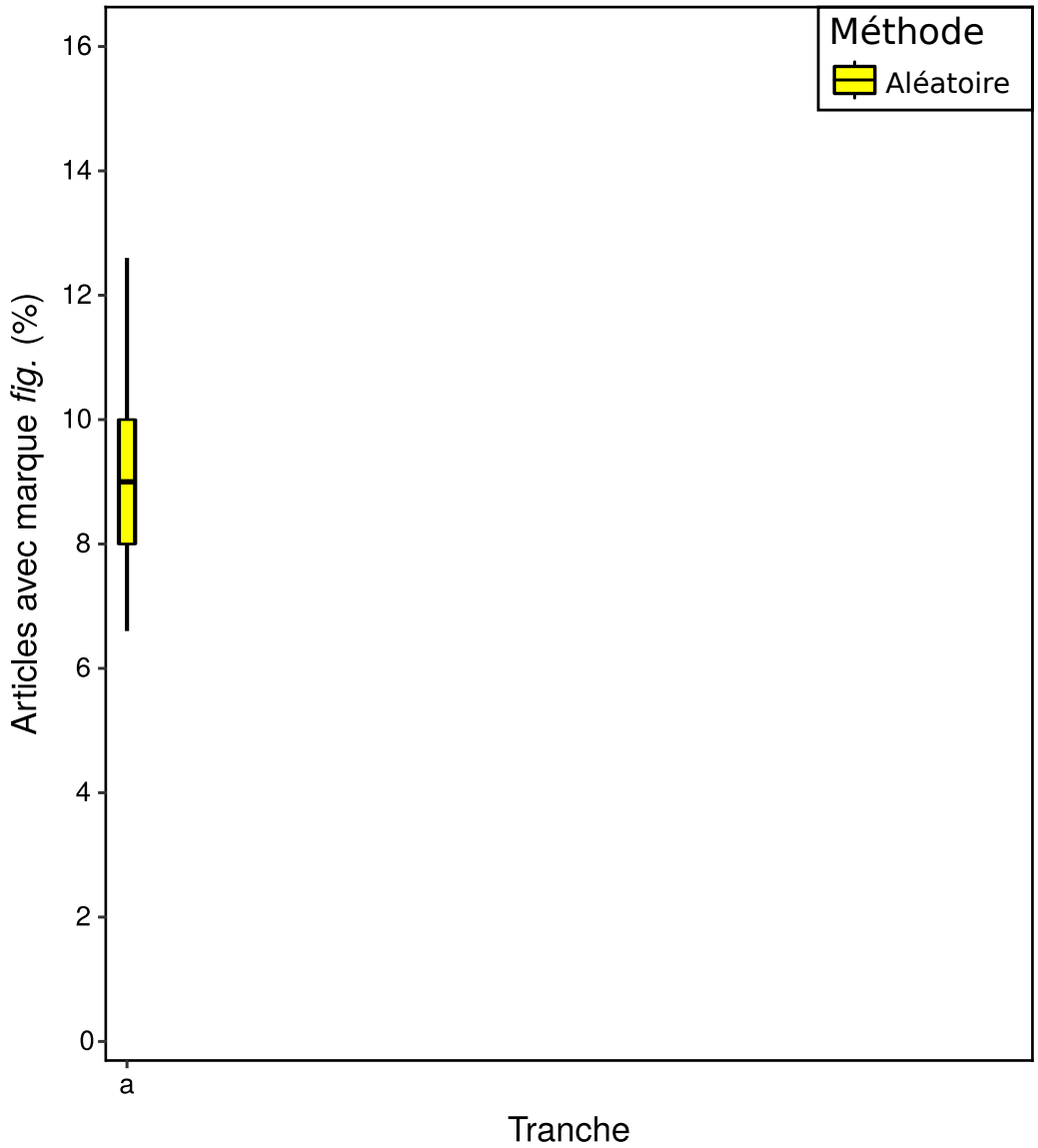
Échantillonnage : zone contiguë *vs.* tirage aléatoire



distribution des valeurs des échantillons

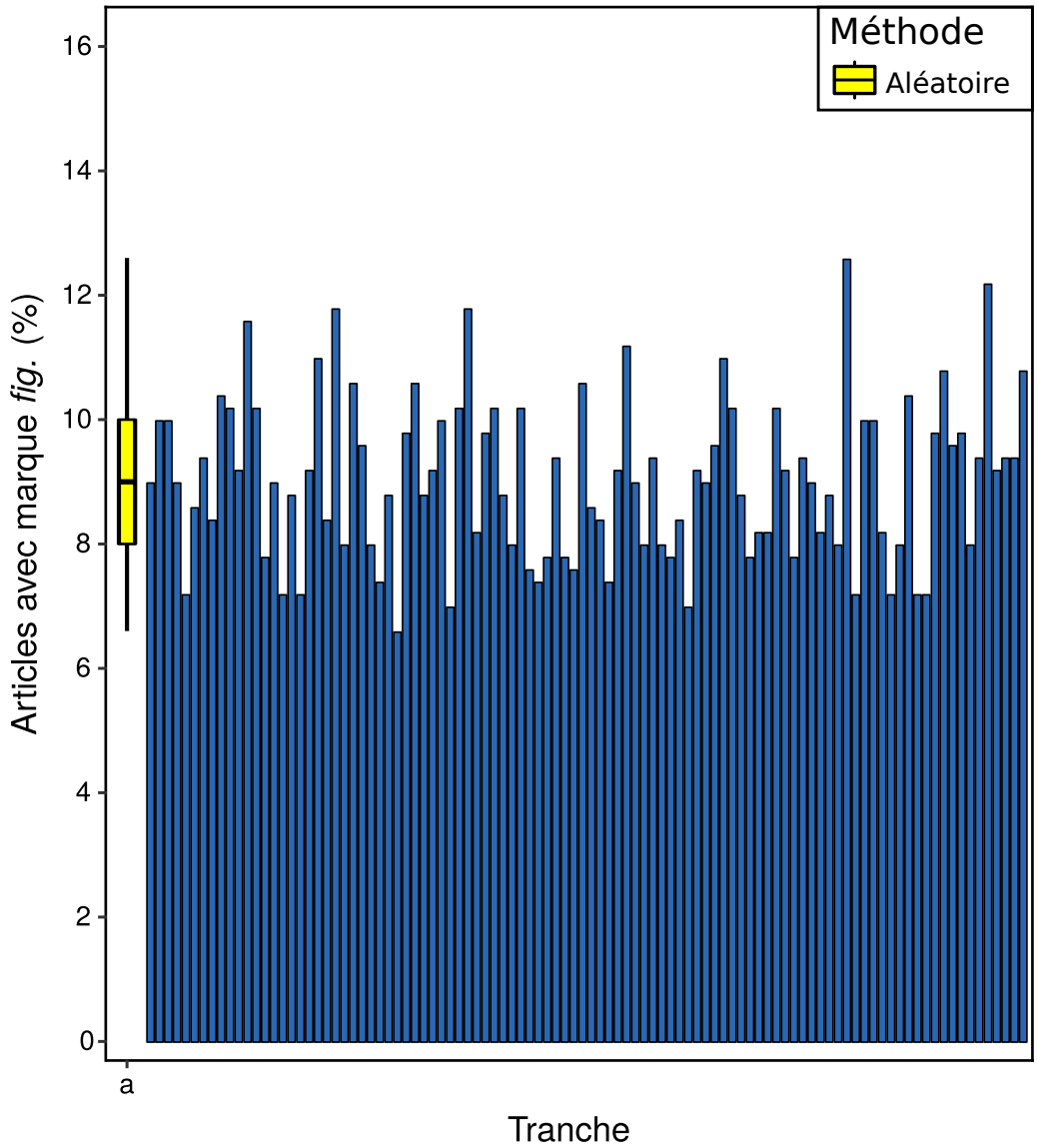
→ 26 tranches (lettres initiales) x 100 échantillons de 500 articles

Échantillonnage : zone contiguë *vs.* tirage aléatoire



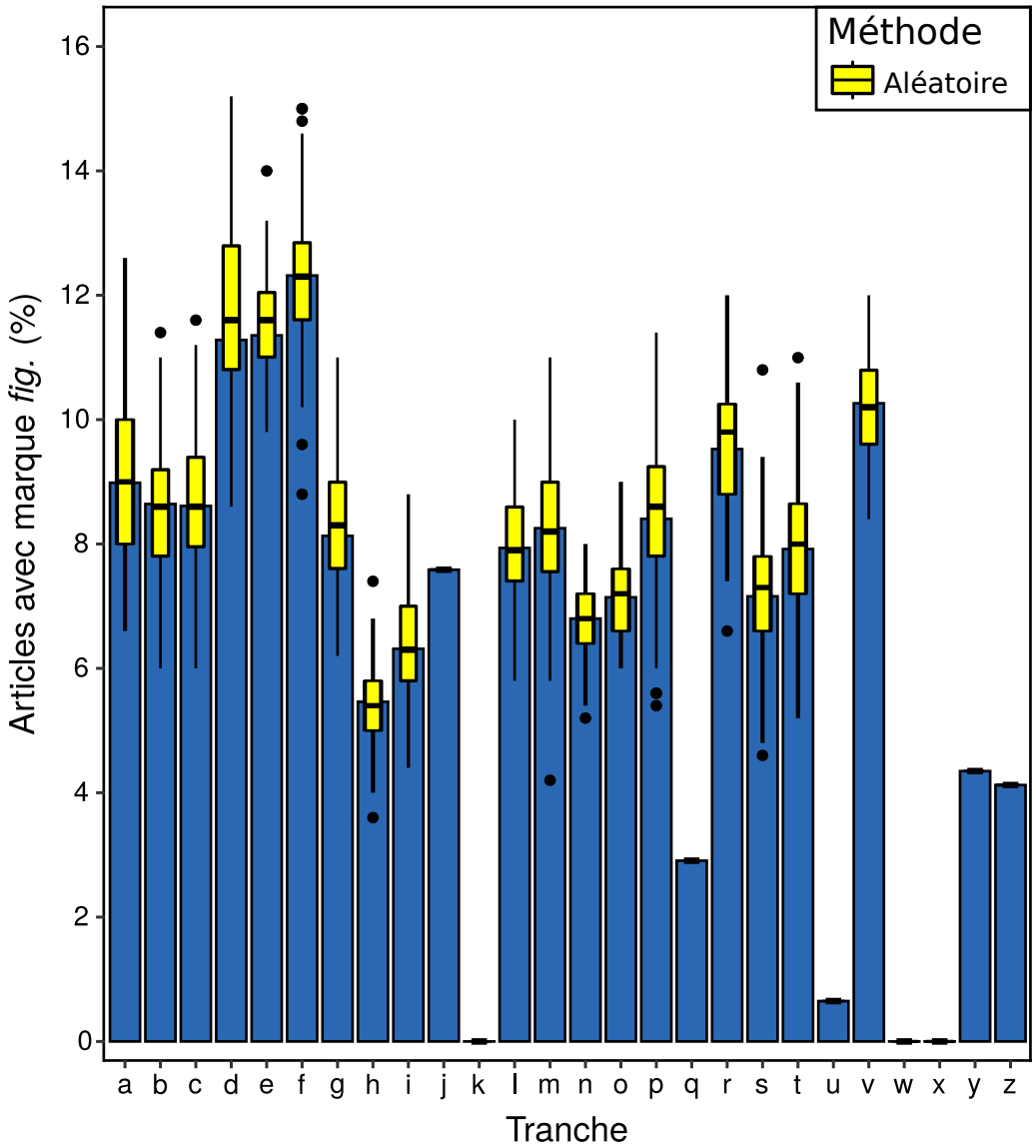
1 boxplot → 100 échantillons de 500 articles d'une tranche donnée

Échantillonnage : zone contiguë *vs.* tirage aléatoire



1 boxplot → 100 échantillons de 500 articles d'une tranche donnée

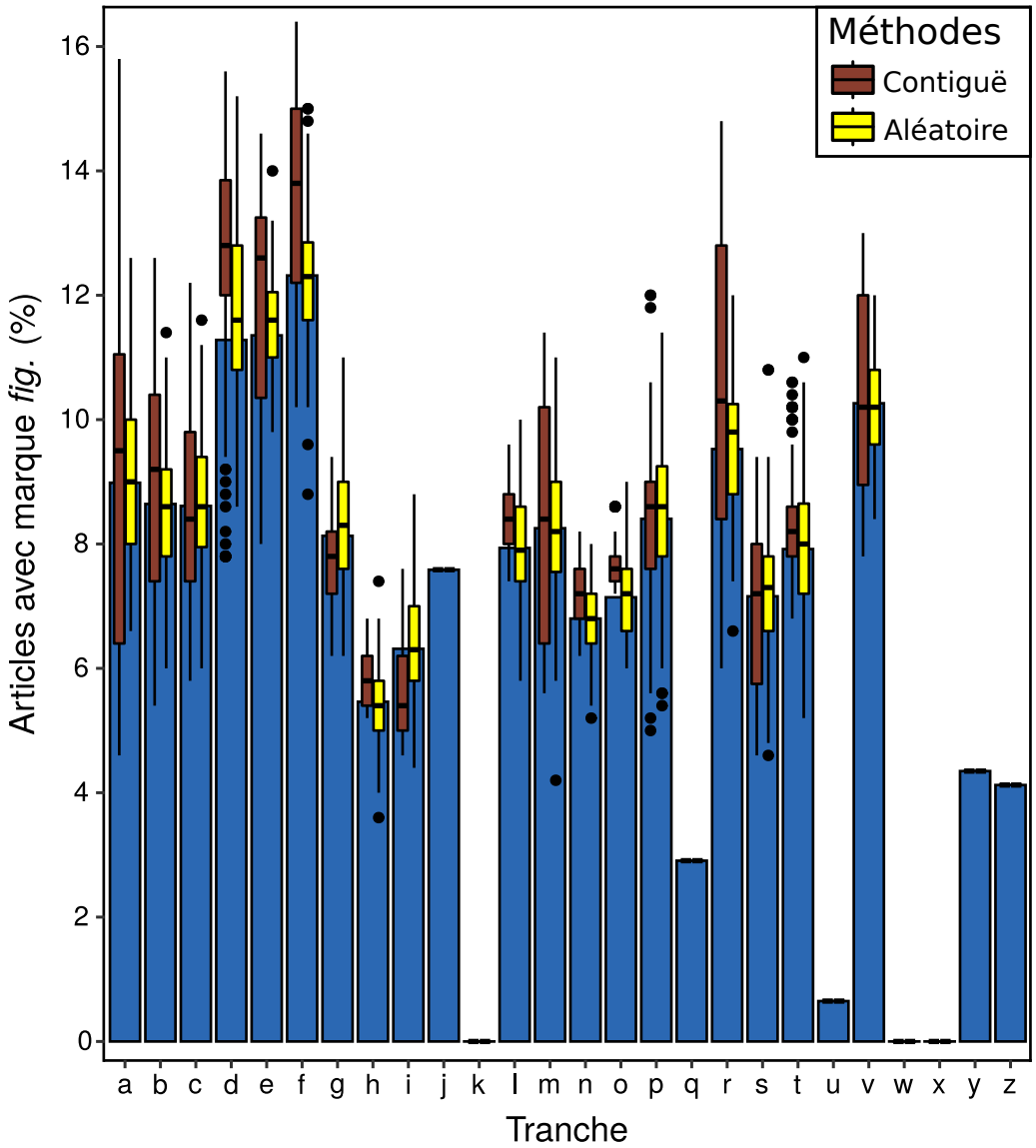
Échantillonnage : zone contiguë *vs.* tirage aléatoire



distribution des valeurs des échantillons

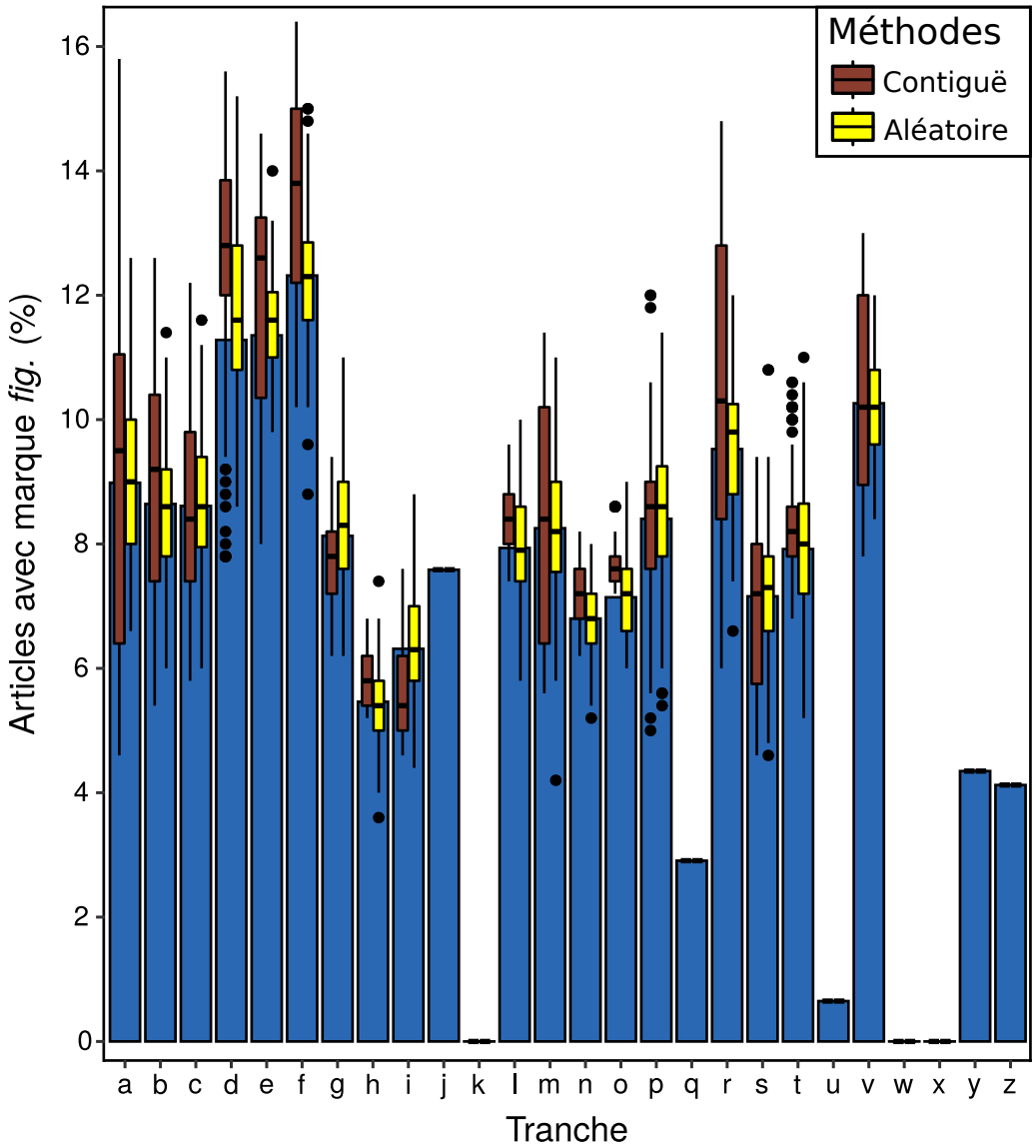
→ 26 tranches (lettres initiales) x 100 échantillons de 500 articles

Échantillonnage : zone contiguë *vs.* tirage aléatoire



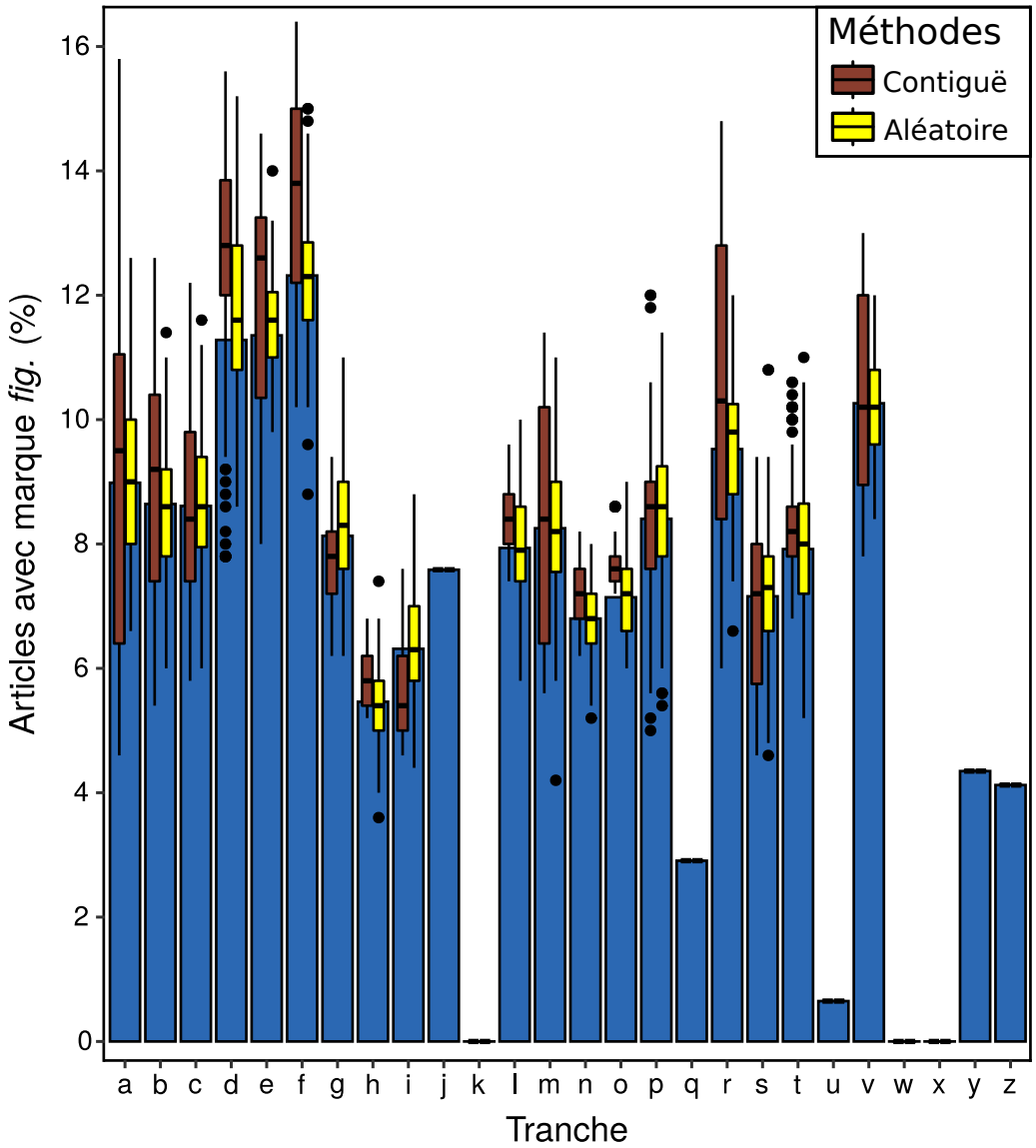
distribution des valeurs des échantillons, par méthode

Échantillonnage : zone contiguë *vs.* tirage aléatoire



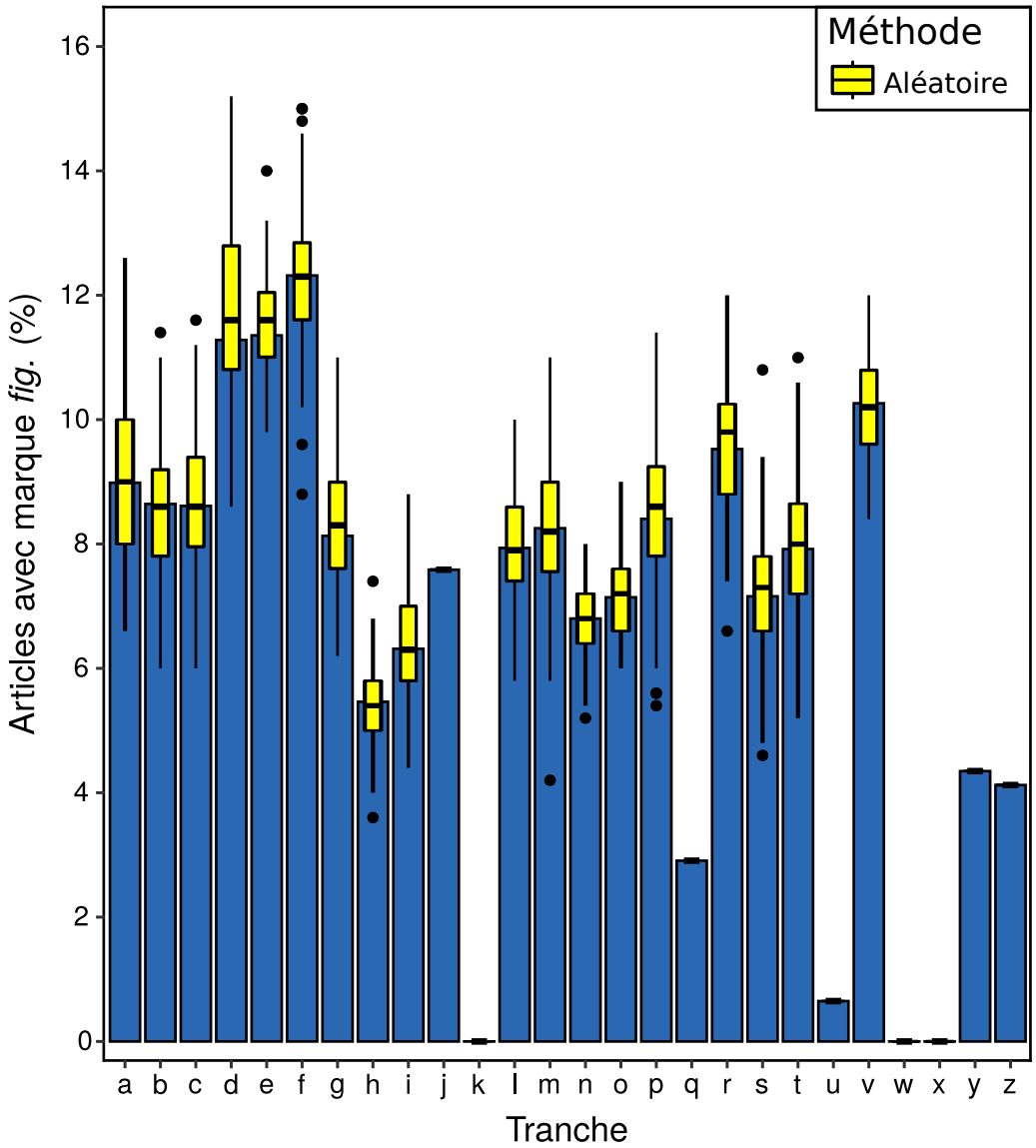
échantillons contigus : plus grande variabilité des distributions

Échantillonnage : zone contiguë *vs.* tirage aléatoire



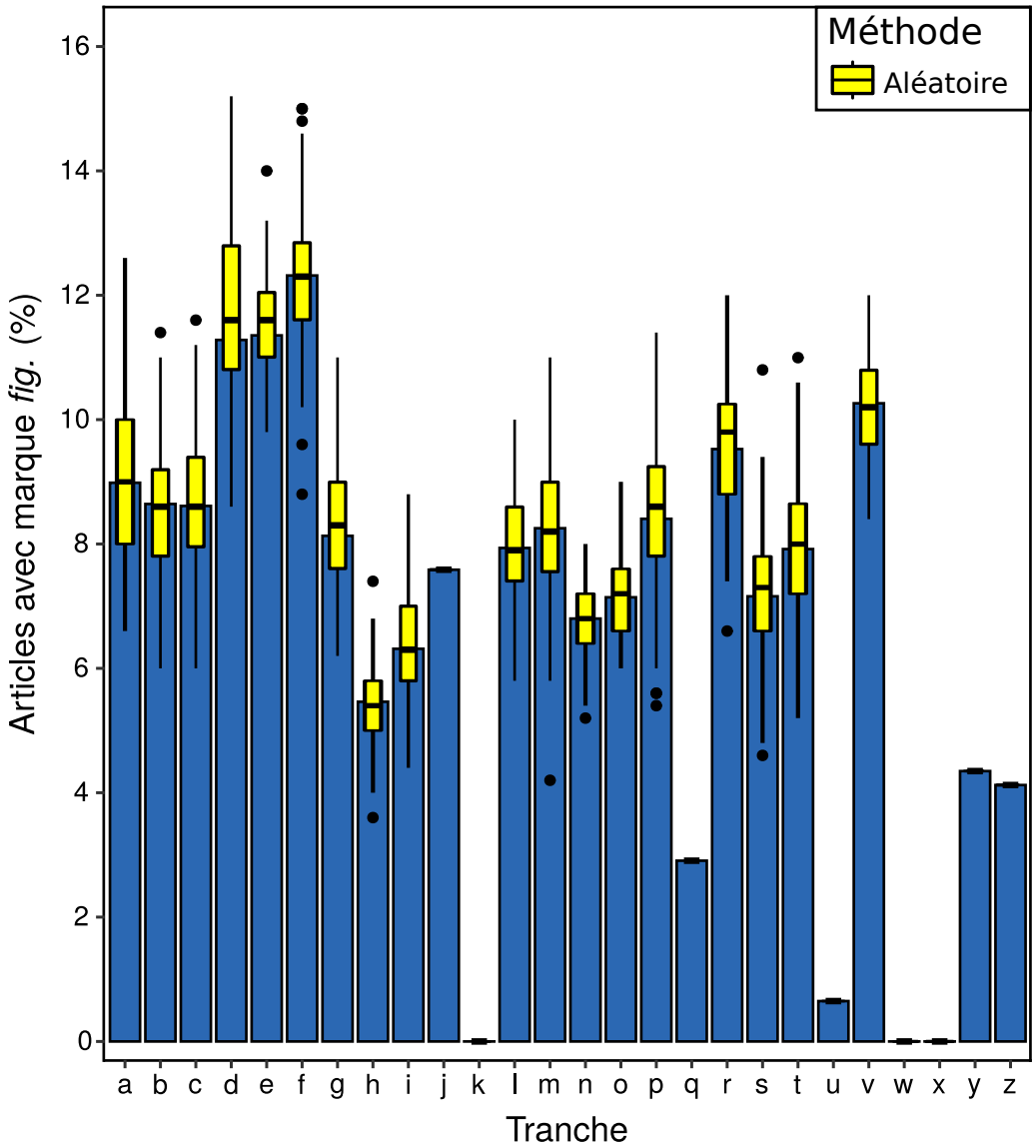
échantillons contigus : distributions moins centrées sur % réel des tranches

Échantillonnage : zone contiguë *vs.* tirage aléatoire



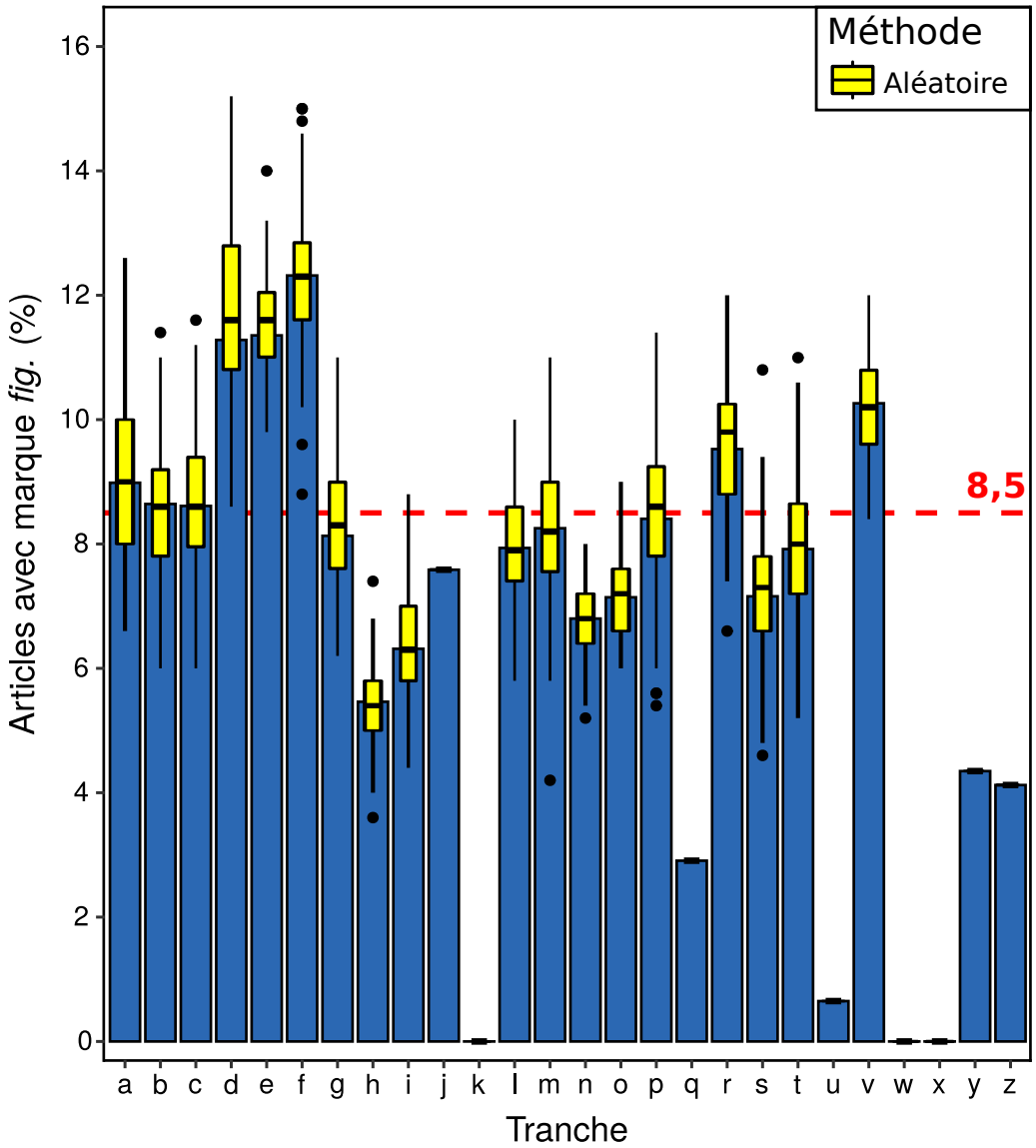
très grande variabilité, même pour la (meilleure) méthode aléatoire

Échantillonnage : zone contiguë *vs.* tirage aléatoire



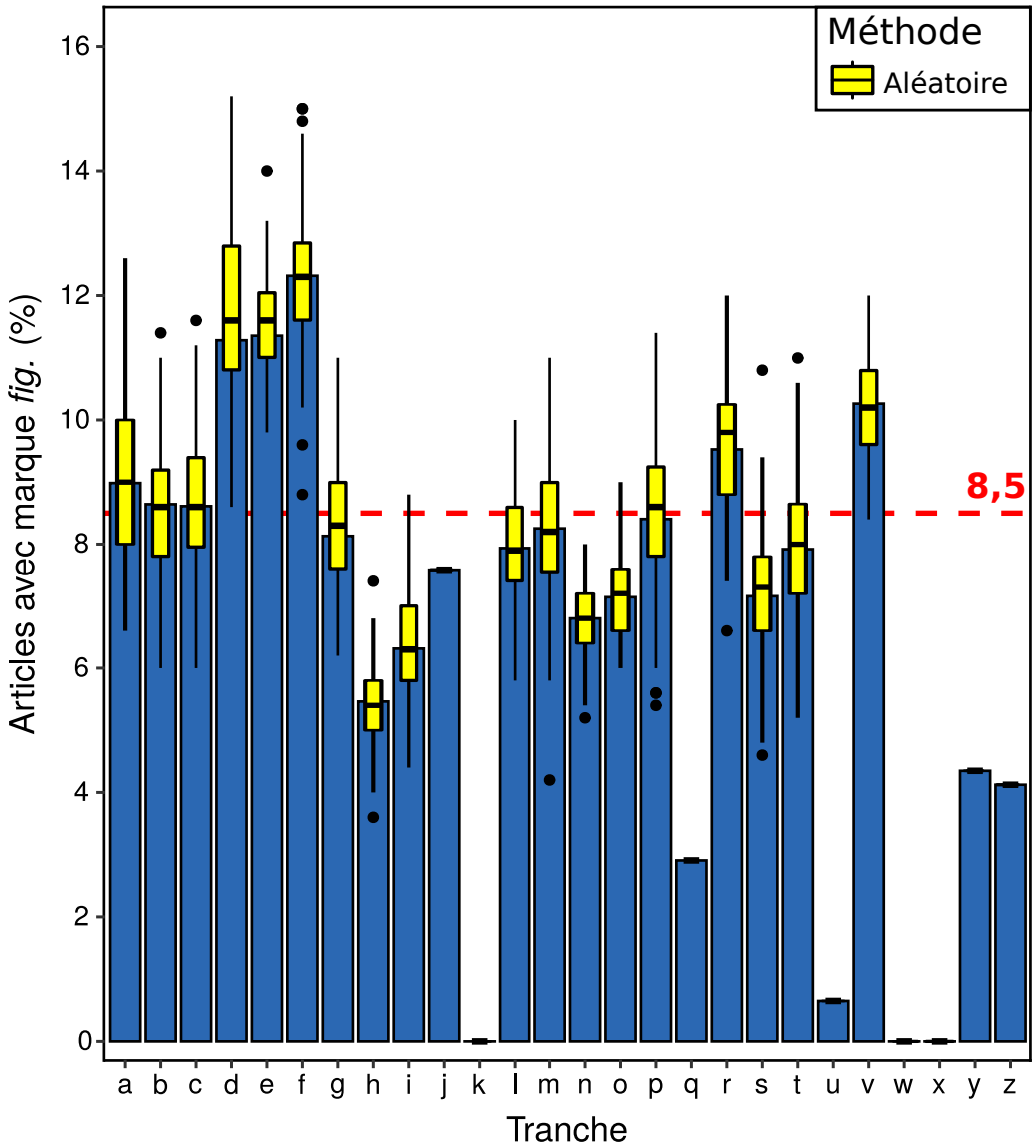
variabilité, même dans les “grosses” tranches : $<4\% \rightarrow >15\%$

Échantillonnage : zone contiguë *vs.* tirage aléatoire



% réel du marquage sur l'ensemble du dictionnaire = 8,5 %

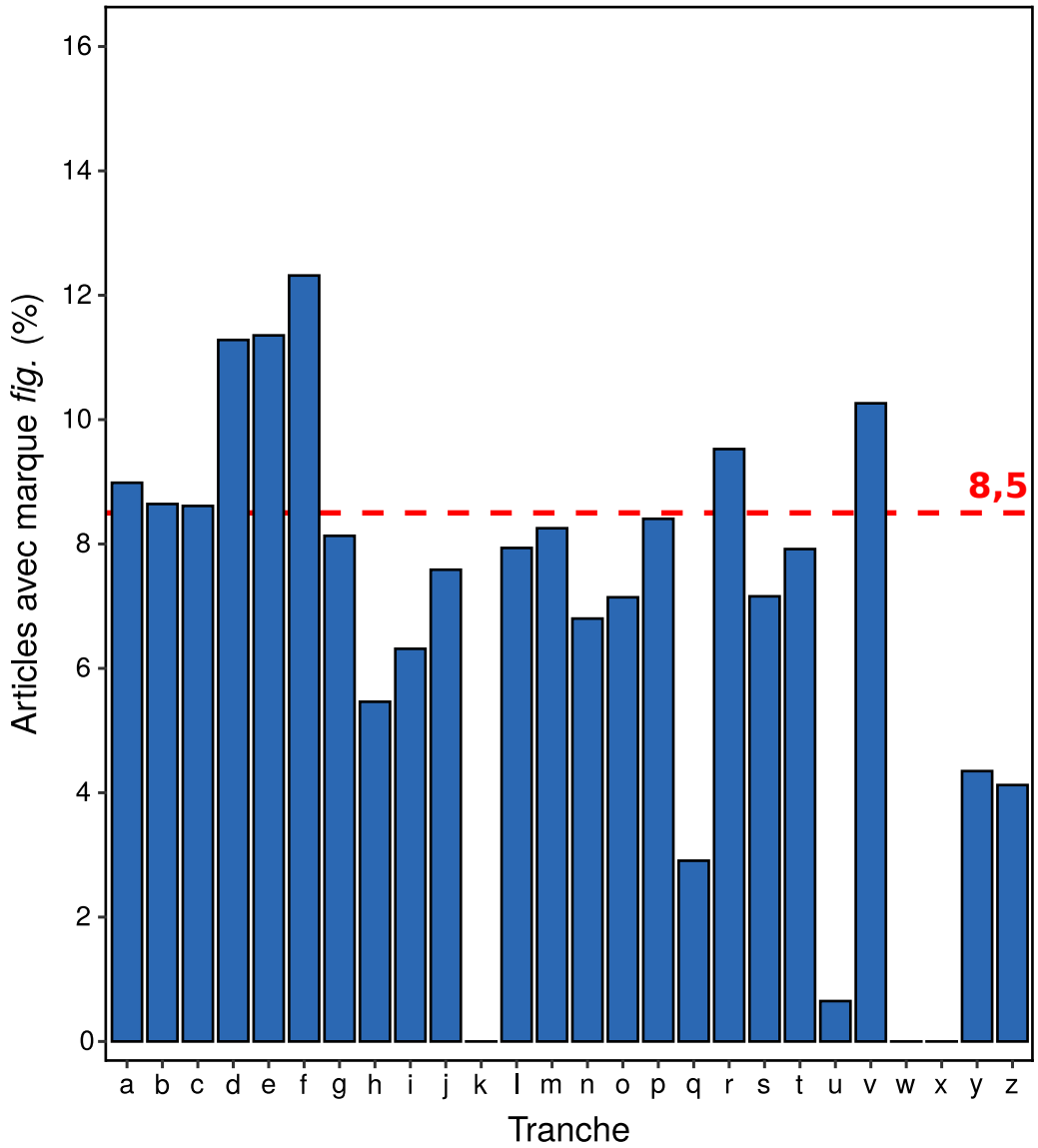
Échantillonnage : zone contiguë *vs.* tirage aléatoire



% réel du marquage sur l'ensemble du dictionnaire = 8,5 %

pour certaines tranches, aucune distribution ne contient cette valeur

Échantillonnage : zone contiguë *vs.* tirage aléatoire



% réel marquage tranches *b*, *c* et *p* ≈ % dictionnaire

Échantillonnage : “bonne” tranche

Choix d’une “bonne” tranche

Indépendamment de la technique d’échantillonnage :

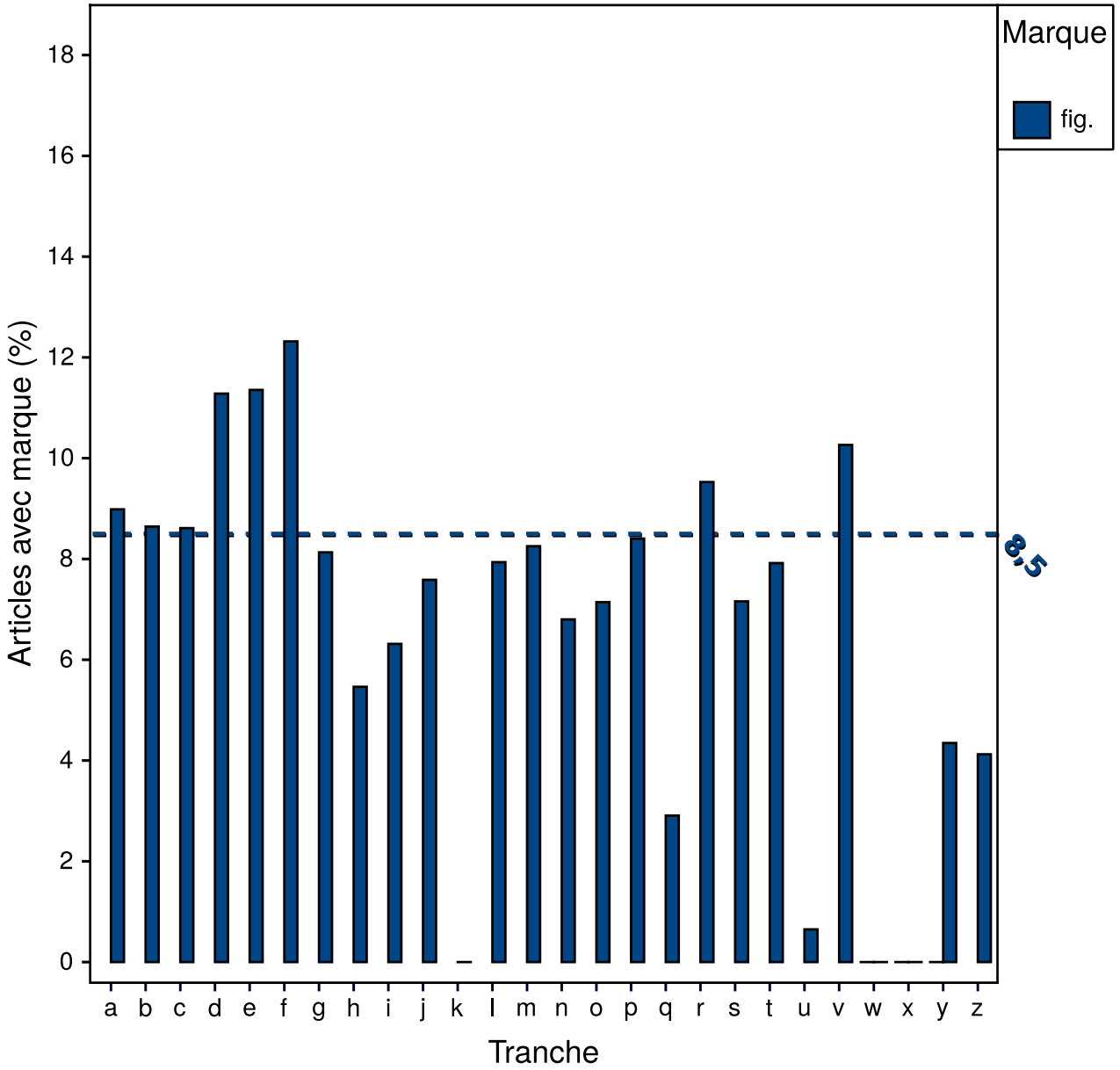
- de quelle tranche tirer un échantillon ?
- ∃ de bonnes tranches *généralement* représentatives ?
  e.g. tranches *b*, *c* et *p*, les plus représentatives pour l’expérience précédente (marque *fig.*) ?

Marque *fam.* : pourcentage d’articles marqués

Sur la totalité des 31 310 noms d’*Usito* :

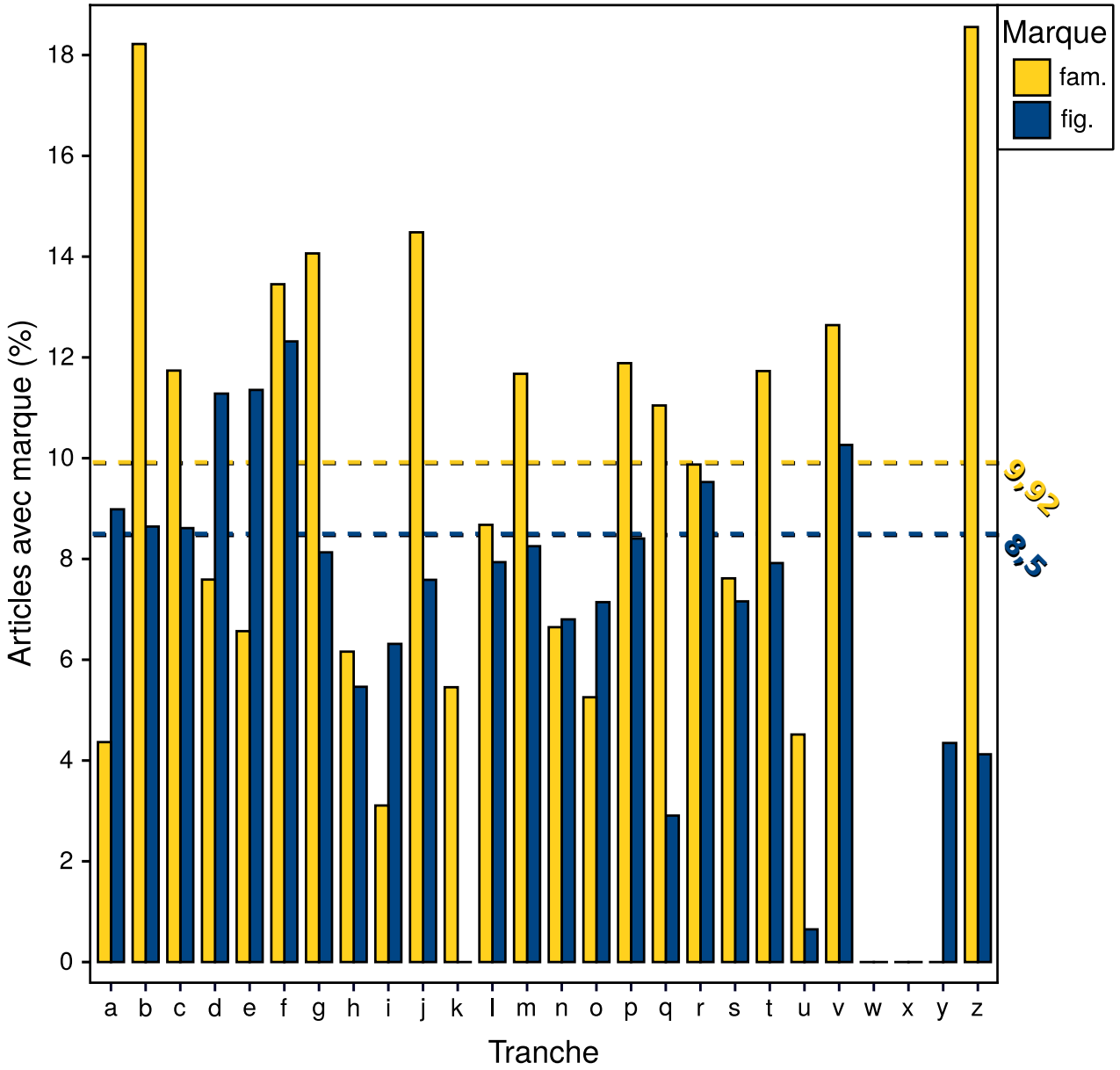
- 3 106 articles contiennent (au moins) une marque *fam.*
- soit 9,92%

Échantillonnage : “bonne” tranche - marques *fig.* et *fam.*



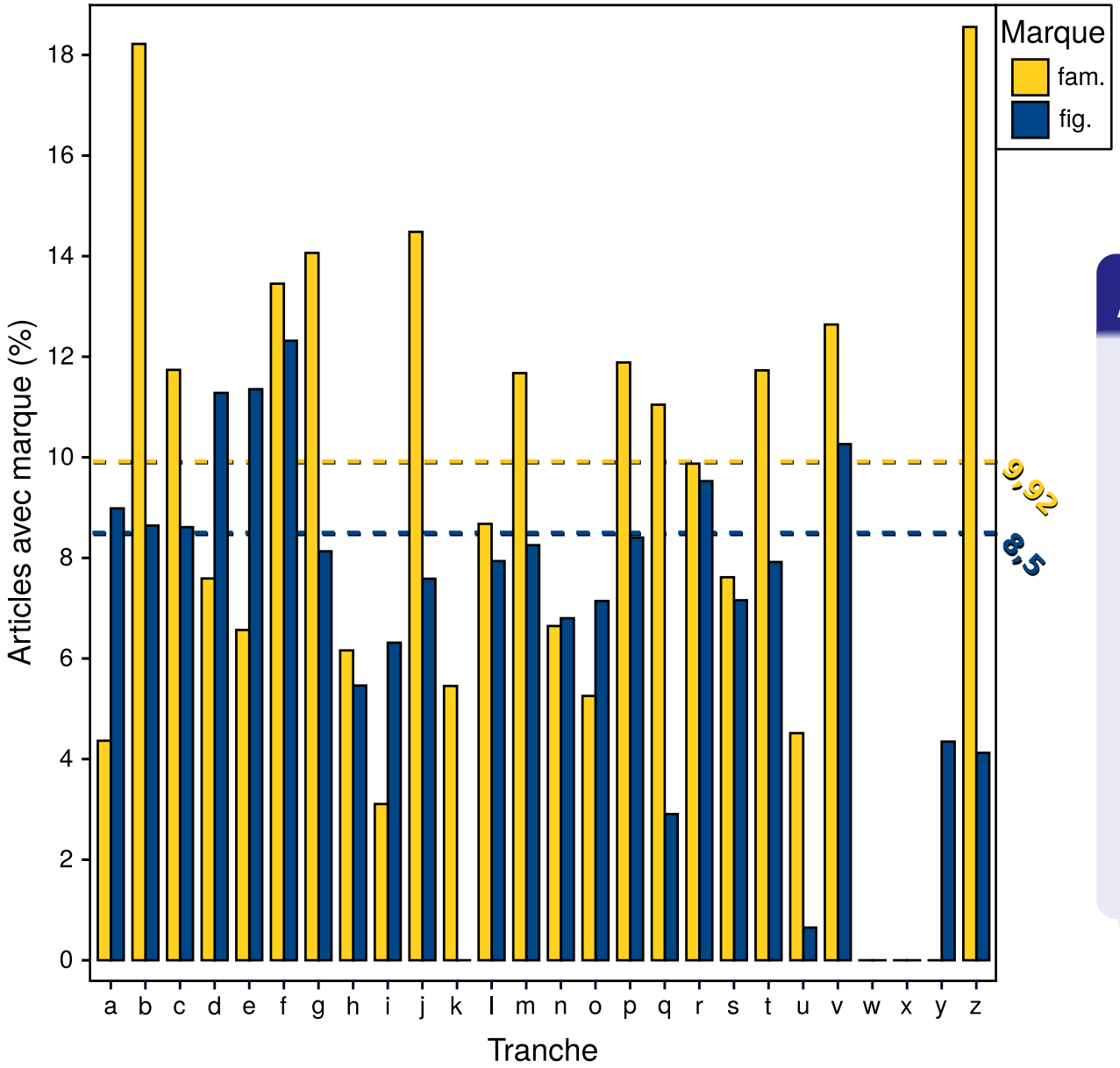
% réel par tranche *vs.* global

Échantillonnage : “bonne” tranche - marques *fig.* et *fam.*



% réel par tranche *vs.* global

Échantillonnage : “bonne” tranche - marques *fig.* et *fam.*



% réel par tranche *vs.* global

À retenir

1. méthode par tranche contiguë à proscrire → méthode aléatoire à privilégier
2. ∄ « bonne tranche » (y compris dans milieu de l'alphabet)
3. 2 échantillons ≠ de tranches ≠ ou = → observations très ≠

Échantillonnage aléatoire : tranche *vs.* dictionnaire entier

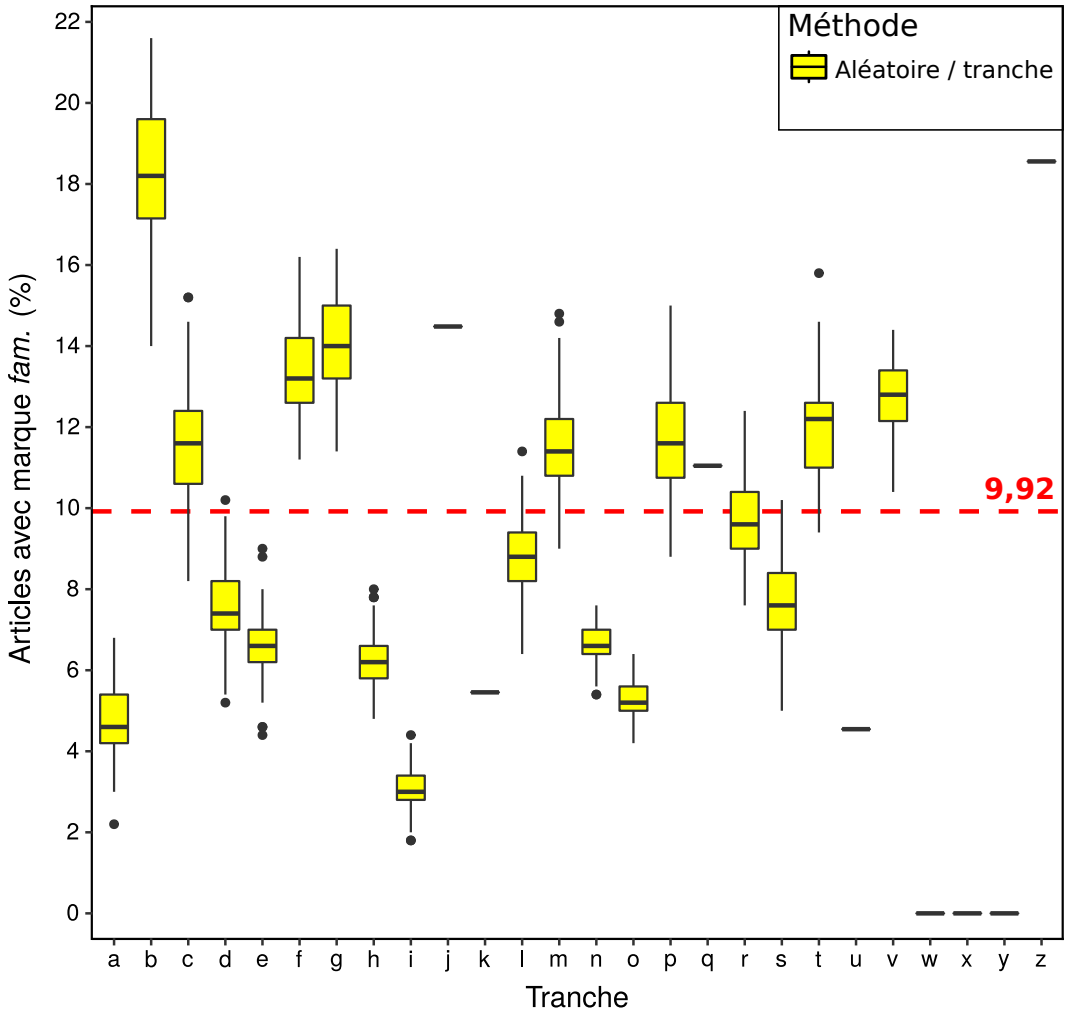
Comparaison échantillons/tranche *vs.* échantillons/dict. global

- 100 échantillons précédents (500 articles tirés aléatoirement par tranche)
- 100 échantillons de 500 articles tirés aléatoirement sur tout le dictionnaire

Échantillonnage aléatoire : tranche *vs.* dictionnaire entier

Comparaison échantillons/tranche *vs.* échantillons/dict. global

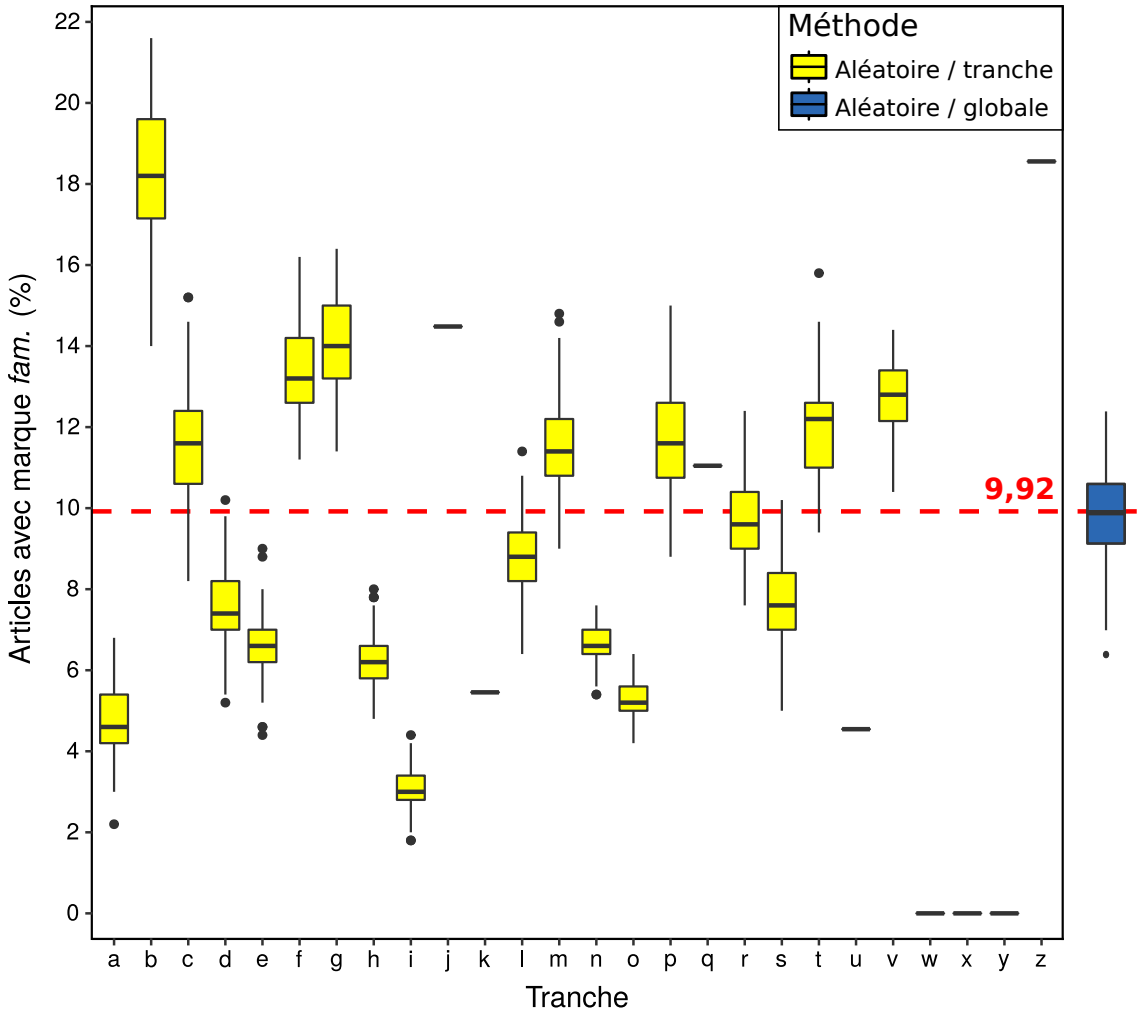
- 100 échantillons précédents (500 articles tirés aléatoirement par tranche)
- 100 échantillons de 500 articles tirés aléatoirement sur tout le dictionnaire



Échantillonnage aléatoire : tranche *vs.* dictionnaire entier

Comparaison échantillons/tranche *vs.* échantillons/dict. global

- 100 échantillons précédents (500 articles tirés aléatoirement par tranche)
- 100 échantillons de 500 articles tirés aléatoirement sur tout le dictionnaire



méthode aléatoire/globale	
min.	6,4 %
médiane	9,9 %
max.	12,4 %

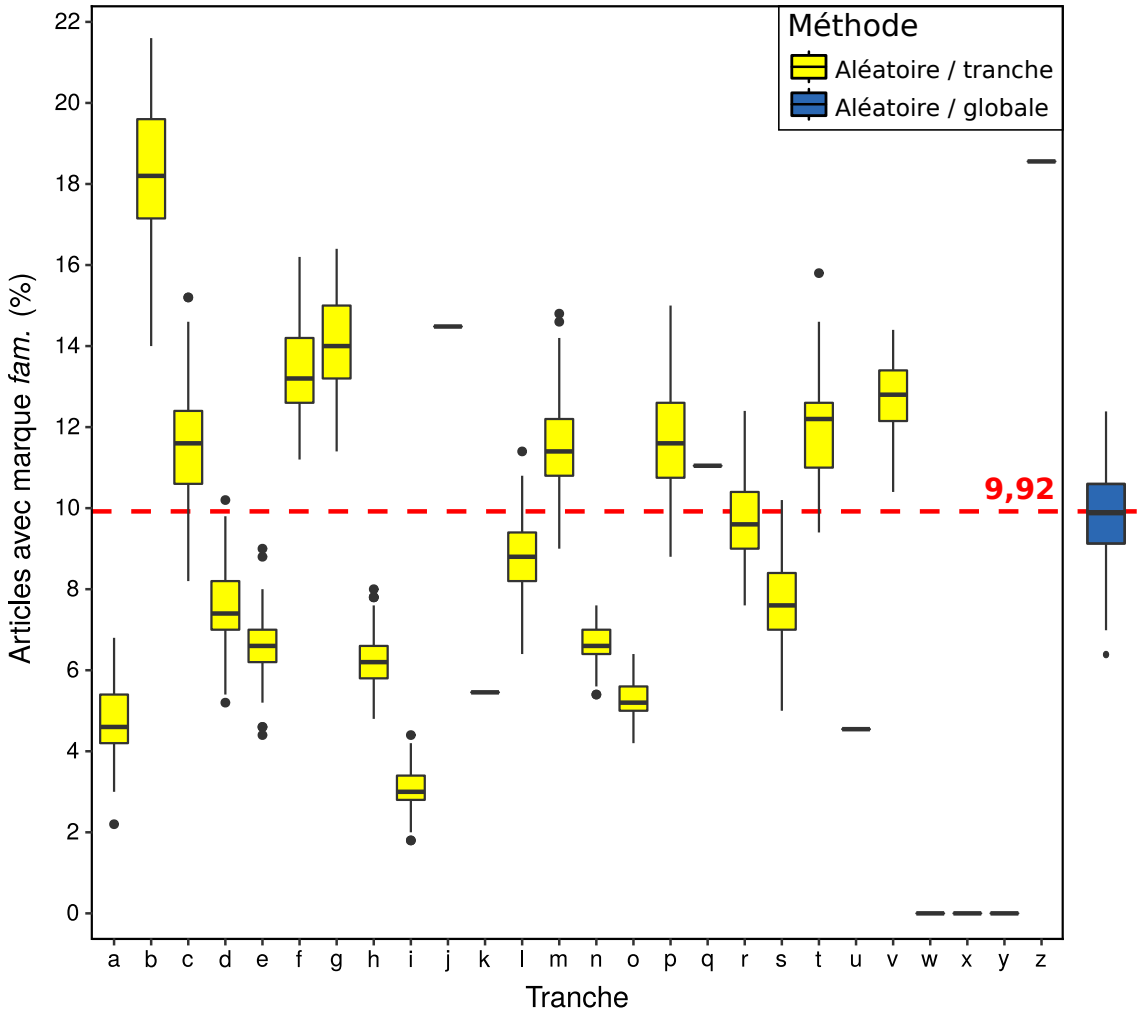
distribution centrée sur valeur réelle mais dispersée

min/max : simple au double

Échantillonnage aléatoire : tranche *vs.* dictionnaire entier

Comparaison échantillons/tranche *vs.* échantillons/dict. global

- 100 échantillons précédents (500 articles tirés aléatoirement par tranche)
- 100 échantillons de 500 articles tirés aléatoirement sur tout le dictionnaire

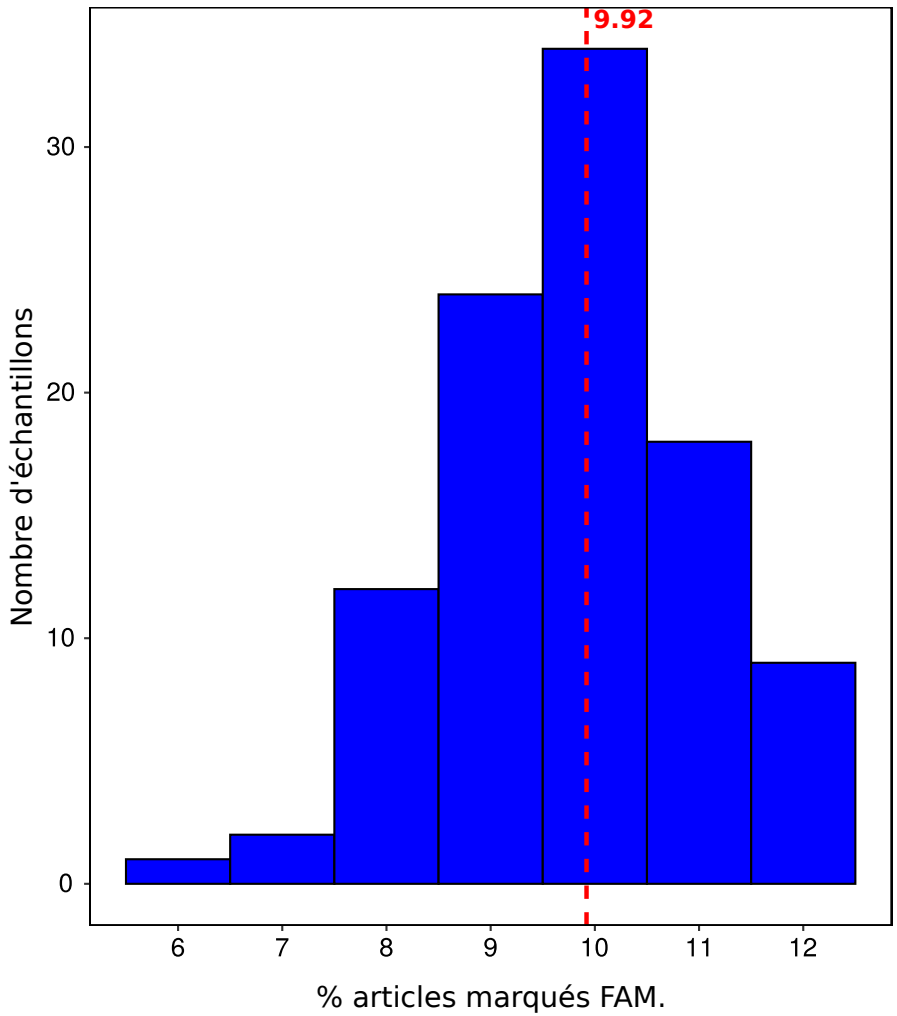


À retenir

1. échantillonnage probabiliste préférable à échantillonnage par zone contiguë
2. échantillonnage sur tout le dictionnaire préférable à échantillonnage dans une tranche donnée

« Contrôler » la représentativité des échantillons

100 échantillons de 500 articles tirés aléatoirement sur tout le dictionnaire



« Contrôler » la représentativité des échantillons

« *Uncontrolled reliability* »

"Most of the samples in current metalexicographic research are judgmental one-stretch samples based on what metalexicographers intuitively consider reliable and representative, usually without having tested this representativeness in any way." (Bukowska, 2010)

Estimation de la fiabilité / représentativité

Pour Bukowska (2010), calcul de la marge d'erreur et (donc) de l'intervalle de confiance.
→ "il y a telle probabilité (= niveau de confiance, *e.g.* 95%) que la véritable proportion p_r soit comprise dans tel intervalle autour de la proportion observée p_o"
$p_r \in [p_o - m_e; p_o + m_e]$ et $m_e = k \times$ erreur type $= k \times \sqrt{\frac{p_o(1-p_o)}{n}}$, avec :

- p_o : proportion observée dans l'échantillon
- p_r : proportion réelle
- n : taille de l'échantillon
- k : coefficient correspondant au niveau de confiance souhaité → table de la loi normale centrée réduite (*e.g.* 1,96 pour un niveau de confiance de 95%)

Expérience précédente : marquage FAM. dans *Usito*

- 100 échantillons de 500 articles, tirés aléatoirement sur tout le dictionnaire (moins pire des méthodes)
- intervalles au niveau de confiance 95% :

p **observée (%)**	**marge d'erreur (%)**	**intervalle de confiance (%)**
6,40 (p_o min)	2,15%	[4,25; 8,55]
9,92 ($p_o = p_r$)	2,62%	[7,3; 12,54]
12,40 (p_o max)	2,89%	[9,51; 15,29]

(au même niveau de confiance, une marge d'erreur de 1% (pour $p_o = 9,92$) nécessite un échantillon de 3500 articles)

« Contrôler » la représentativité des échantillons

Expérience précédente : marquage FAM. dans *Usito*

- 100 échantillons de 500 articles, tirés aléatoirement sur tout le dictionnaire (moins pire des méthodes)
- intervalles au niveau de confiance 95% :

p **observée (%)**	**marge d'erreur (%)**	**intervalle de confiance (%)**
6,40 (p_o min)	2,15%	[4,25; 8,55]
9,92 ($p_o = p_r$)	2,62%	[7,3; 12,54]
12,40 (p_o max)	2,89%	[9,51; 15,29]

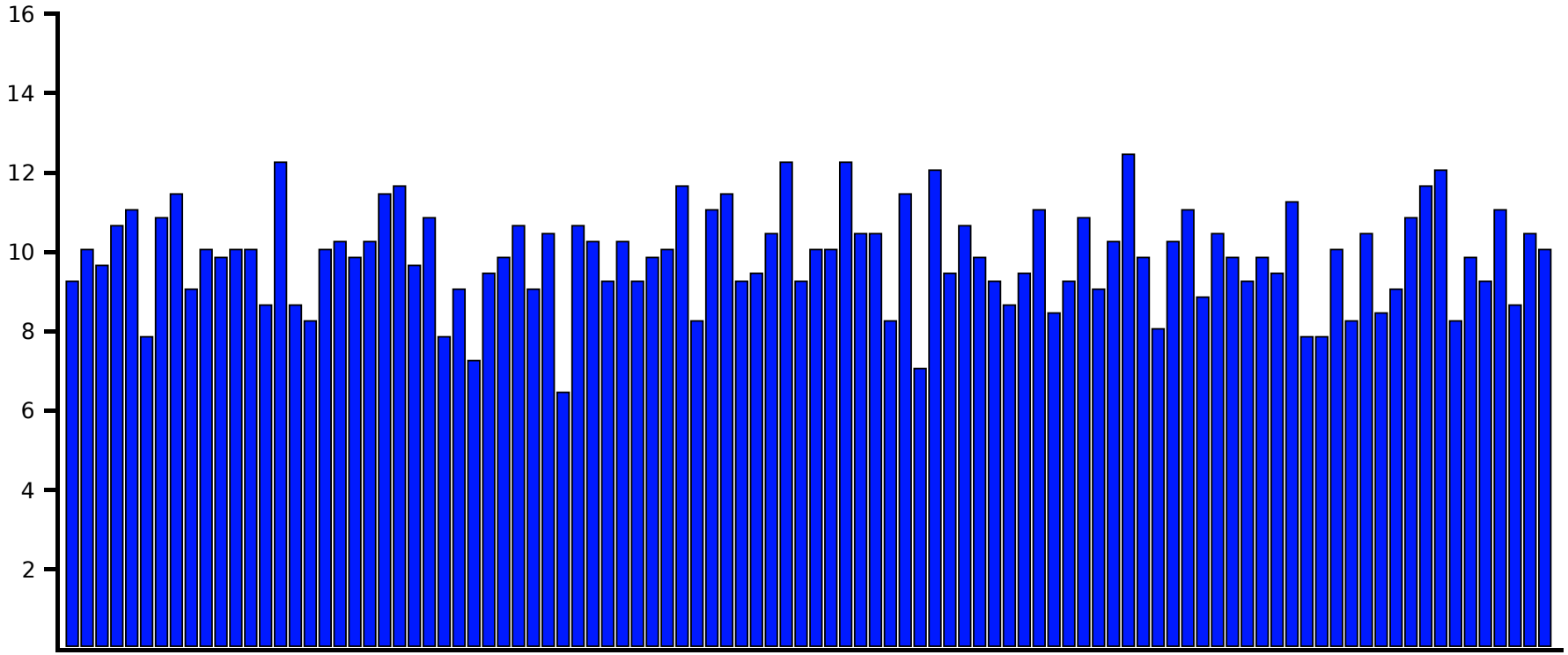
(au même niveau de confiance, une marge d'erreur de 1% (pour $p_o = 9,92$) nécessite un échantillon de 3500 articles)

- marge d'erreur satisfaisante ? Niveau de confiance satisfaisant ?

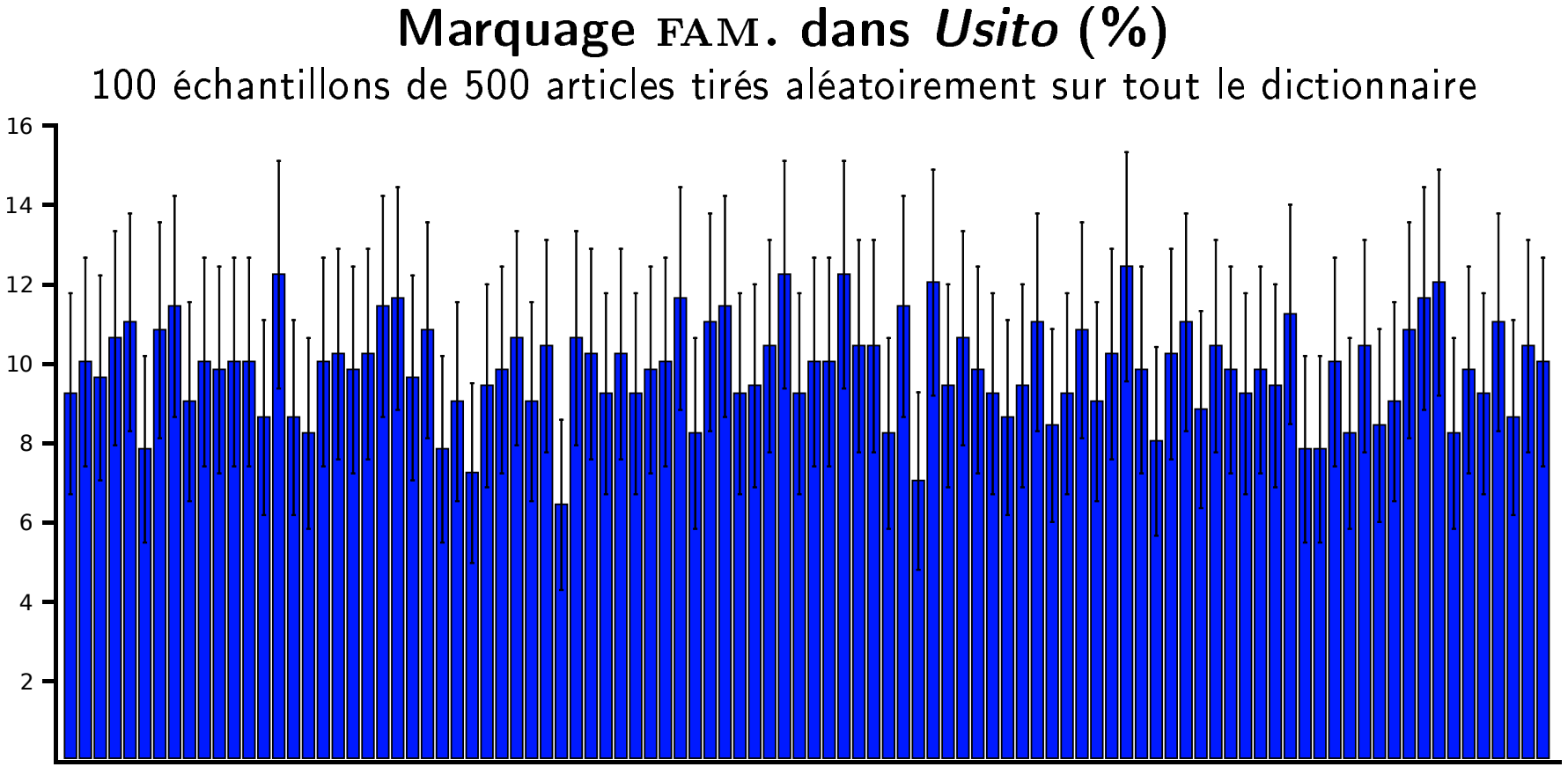
« Contrôler » la représentativité des échantillons

Marquage FAM. dans *Usito* (%)

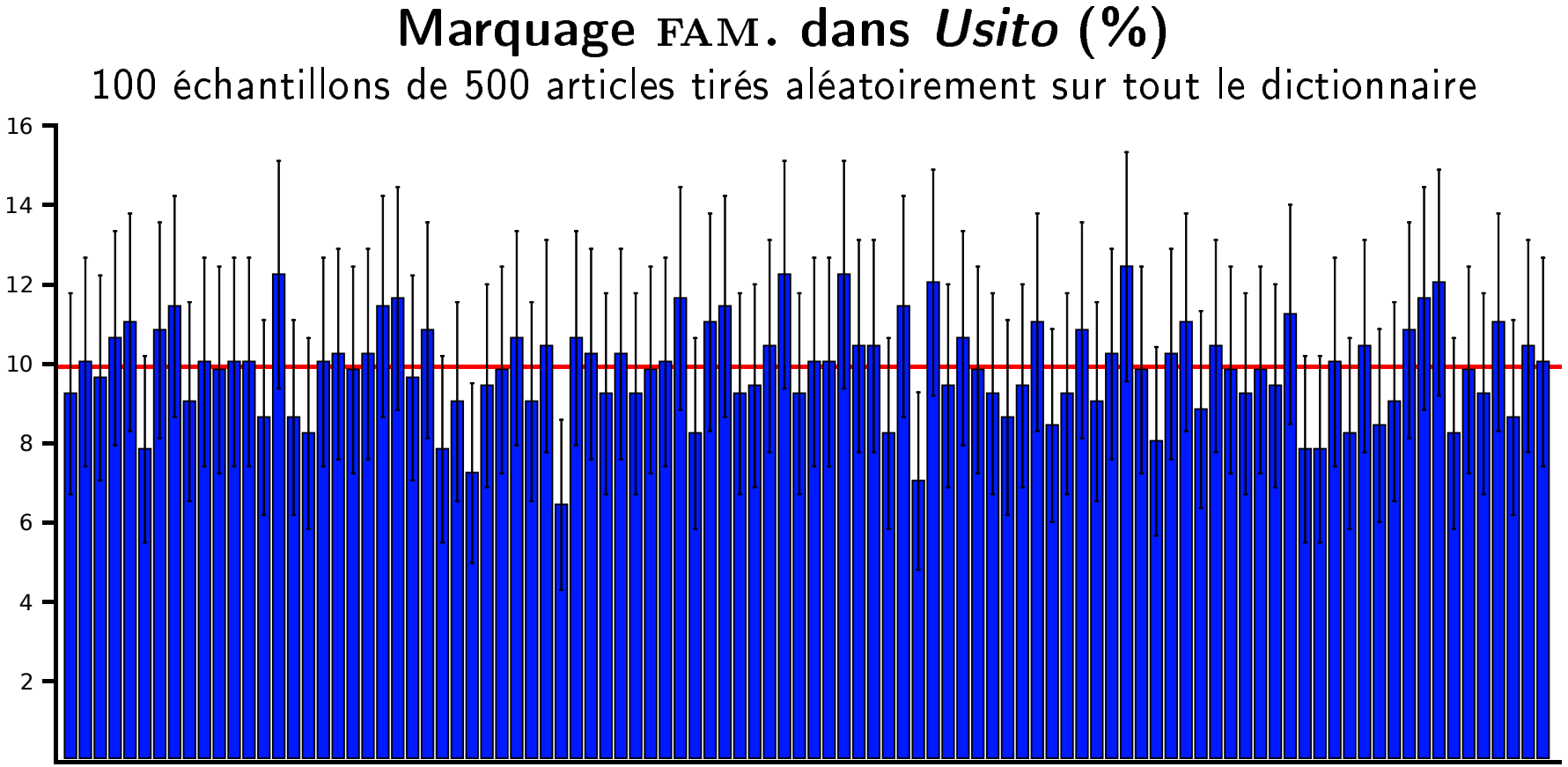
100 échantillons de 500 articles tirés aléatoirement sur tout le dictionnaire



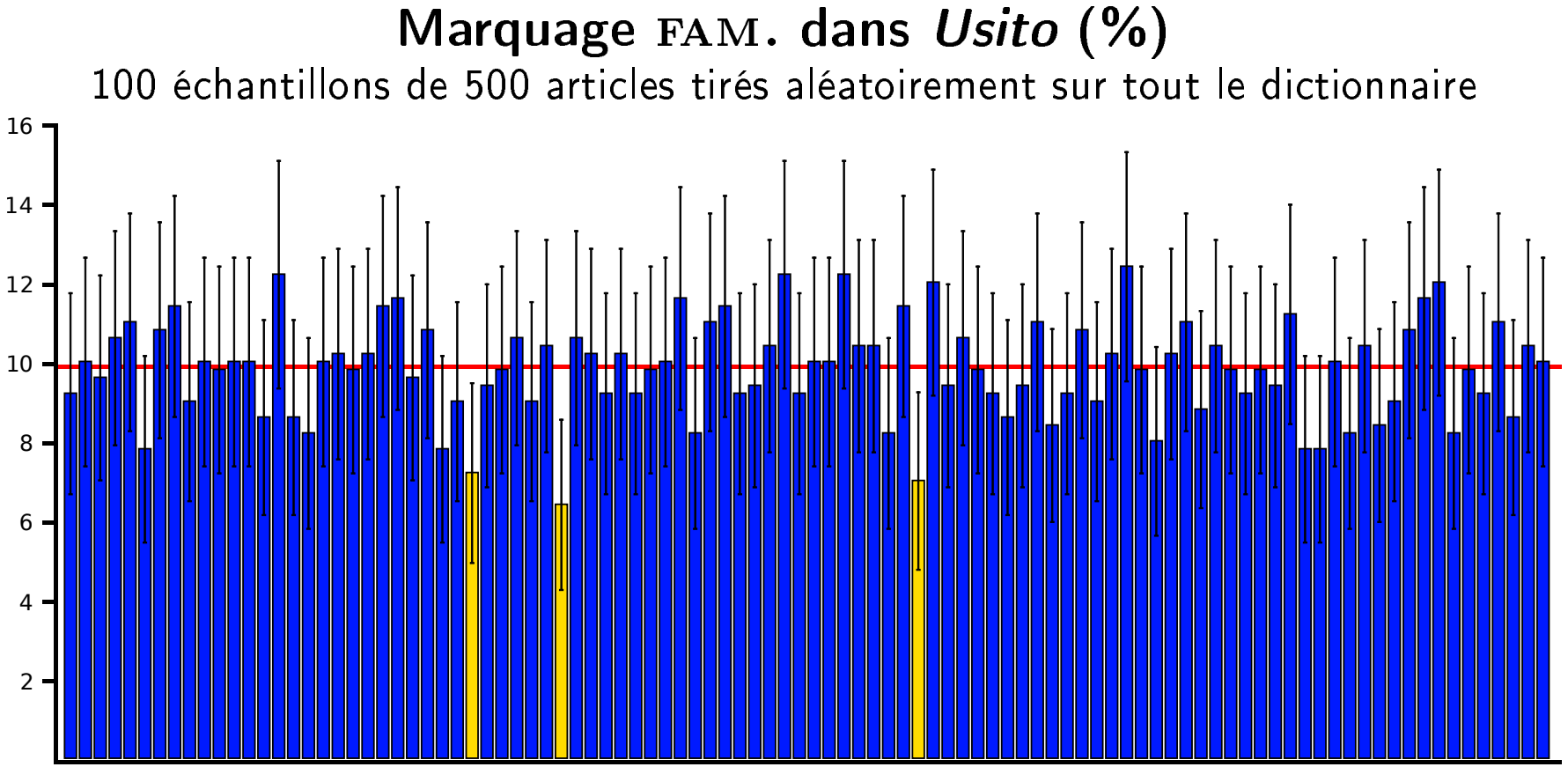
« Contrôler » la représentativité des échantillons



« Contrôler » la représentativité des échantillons



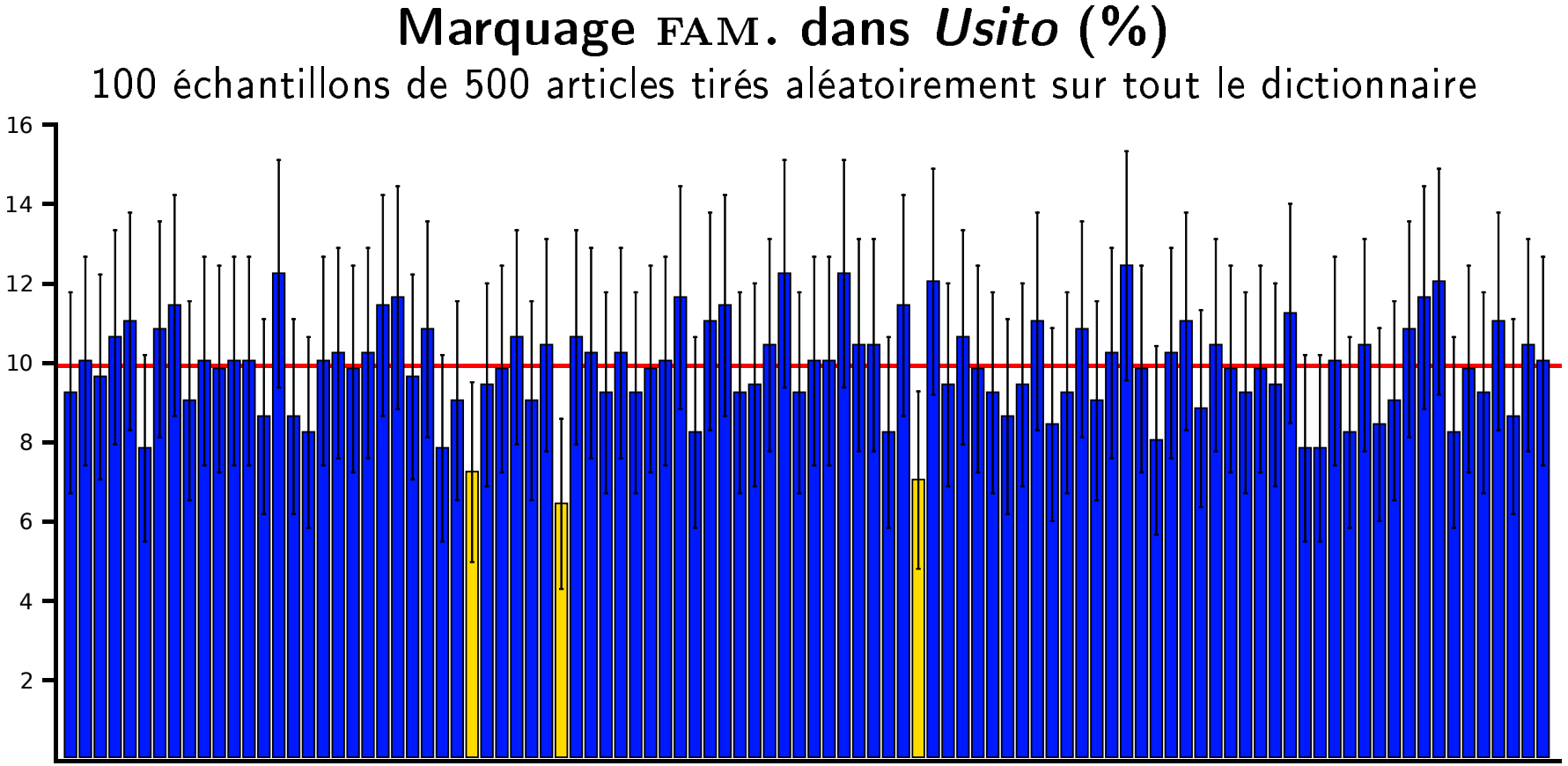
« Contrôler » la représentativité des échantillons



3 échantillons sur 100 pour lesquels la valeur réelle est hors de l'intervalle

(au niveau de confiance 95%)

« Contrôler » la représentativité des échantillons



3 échantillons sur 100 pour lesquels la valeur réelle est hors de l'intervalle (au niveau de confiance 95%)

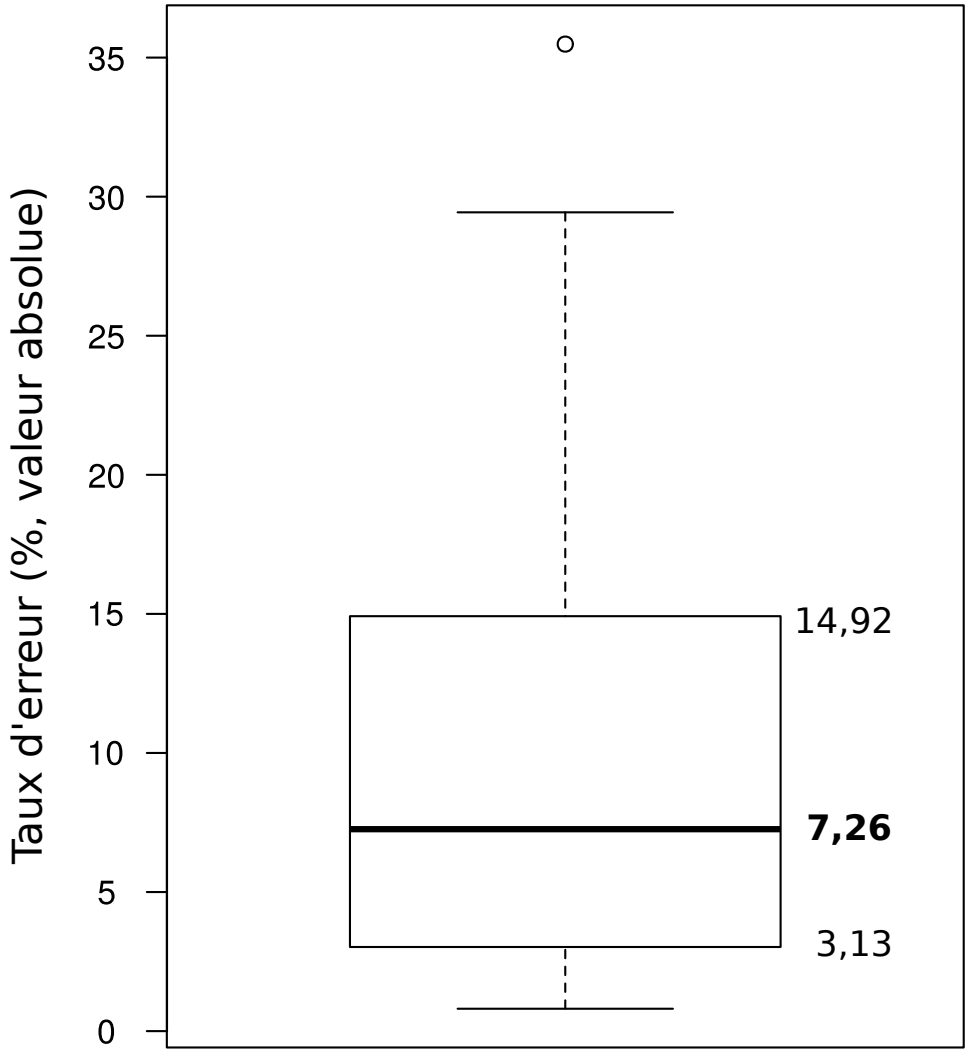
Problème (en plus de la taille de l'intervalle) : l'analyste ne sait pas si son (unique) échantillon est malchanceux

« Contrôler » la représentativité des échantillons

Taux d'erreur

Pour un échantillon :

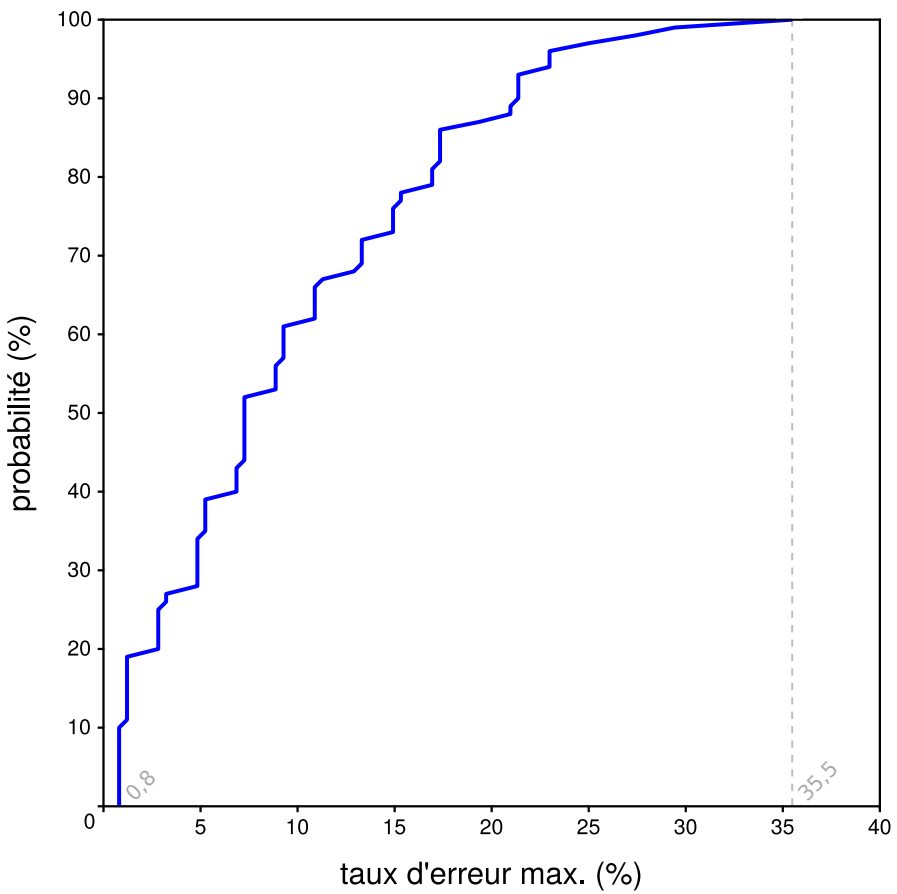
$$T_e = \frac{|p_o - p_r|}{p_r}$$



Distribution du taux d'erreur pour les 100 échantillons (tirage aléatoire sur tout le dictionnaire)

« Contrôler » la représentativité des échantillons

Proportion d'échantillons (parmi les 100 générés), d'afficher un taux d'erreur inférieur/supérieur à un seuil donné

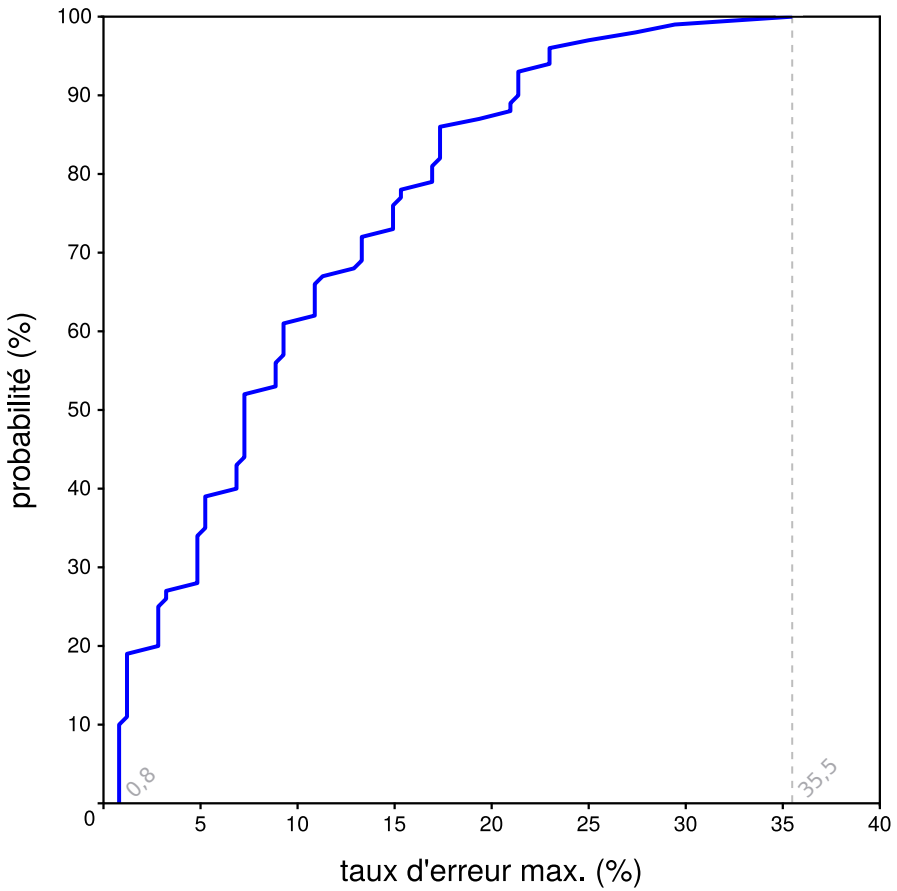


Un échantillon a :

- 10% de chances d'afficher un taux d'erreur inférieur à 1%
- 19% de chances d'afficher un taux d'erreur inférieur à 2%
- 34% de chances d'afficher un taux d'erreur inférieur à 5%
- 61% de chances d'afficher un taux d'erreur inférieur à 10%
- 87% de chances d'afficher un taux d'erreur inférieur à 20%
- 100% de chances d'afficher un taux d'erreur inférieur à 36%

« Contrôler » la représentativité des échantillons

Proportion d'échantillons (parmi les 100 générés), d'afficher un taux d'erreur inférieur/supérieur à un seuil donné



Un échantillon a :

- 90% de chances d'afficher un taux d'erreur supérieur à 1%
- 81% de chances d'afficher un taux d'erreur supérieur à 2%
- 66% de chances d'afficher un taux d'erreur supérieur à 5%
- 39% de chances d'afficher un taux d'erreur supérieur à 10%
- 13% de chances d'afficher un taux d'erreur supérieur à 20%
- 0% de chances d'afficher un taux d'erreur supérieur à 36%

Analyse par échantillonnage : conclusions (1)

À retenir

1. échantillonnage probabiliste préférable à échantillonnage par zone contiguë
2. échantillonnage sur tout le dictionnaire préférable à échantillonnage dans une tranche donnée
3. aucune garantie satisfaisante de la représentativité d'un échantillon

III - EXPÉRIENCES D'ÉCHANTILLONNAGE

Type d'étude / dictionnaire analysé

1. Étude synchronique : *Usito*
2. **Étude diachronique : tomes du TLF**

Projets au long cours : variations des pratiques éditoriales

Alphabet fatigue (Osselton, 2007)

- avant l'informatique, travail sous "la tyrannie de l'alphabet, de A à Z"
- traitement plus fouillé au début qu'à la fin
- dictionnaires anglais actuels (1995-2004), mot médian = *litteral*, *lotto*, *Lycra*, *machinable*, *main*, *market*
- dictionnaires du XVII^{e} s. : entre les tranches *hu-* et *lo-*
- explication du déséquilibre : pression temporelle et financière des maisons d'édition et des imprimeurs ($\rightarrow$ coupe dans la nomenclature, longueur des articles), mais également nombreux autres facteurs explicatifs !
- déséquilibres inverses également observés

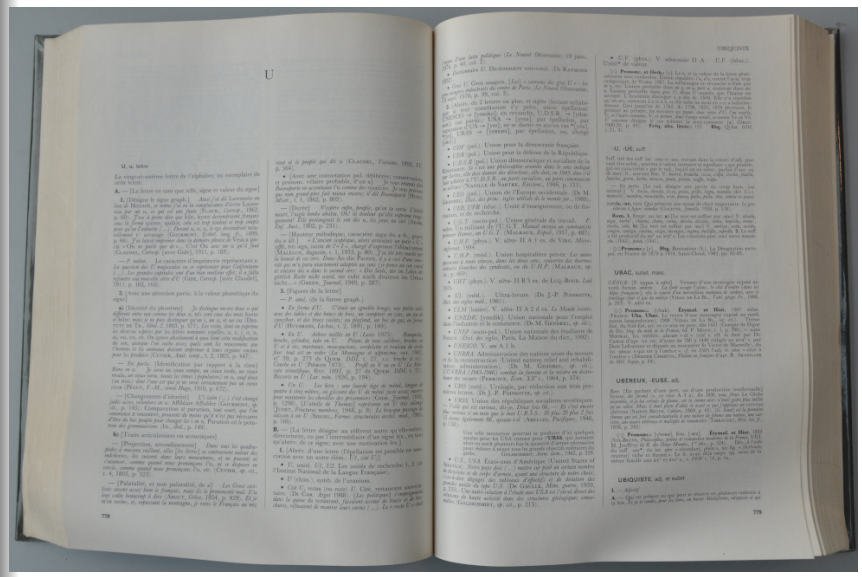
Forensic dictionary analysis (Coleman & Ogilvie, 2009)

étude de faits dictionnairiques $\rightarrow$ indices (et conséquences) du processus de conception
(qui contredisent parfois le paratexte et autres communications des éditeurs)

Le *Trésor de la langue française*

Le projet / le produit

1. dictionnaire institutionnel, général, monolingue, « de langue »
2. conception fondée sur l'exploitation d'un corpus essentiellement littéraire (surtout au début)
3. 16 tomes (consultables à la BUC)
4. publication papier : 1971-1994 (publication/vente de chaque tome publié au fur et à mesure)
5. numérisation, mise en ligne en 2002



Nombreuses évolutions

Au cours du projet, nombreux changements (*e.g.* de direction)
→ conséquences réelles ou fantasmées sur les tomes successifs

Changements effectifs (Radermacher, 2004)

Changements au niveau de la typographie, de la microstructure, du corpus, du traitement...

- tome III : disparition de la rubrique STYL[istique]
  (connotation possibles d'un mot)
- lettre F (tome VIII) : *élém[ents] préf[ixaux]/suff[ixaux]* → *élém[ents] de compos[ition]*
  lettre G (tome IX) : apparition de *élém[ent] form[ant]*
  (catégories non documentées)
- dès premiers tomes, réduction nb entrées principales de la nomenclature : dérivés → sous-entrées (= « entrées-cachées ») dans les rubriques DÉR[ivés] et REM[arque]
- typographie fluctuante au fil des tomes : taille de caractères « généreuse » et constante dans tome I, réservée à certaines entrées seulement dans tome XVI
- lettres A à C : nomenclature extraite exclusivement du corpus littéraire
  puis ajout presse, documents techniques, autres dictionnaires

Le *Trésor de la langue française* : évolution au fil des tomes

Changements : les exemples (Radermacher, 2004, 2005)

- P. Imbs annonce dès le « *Au lecteur* » du tome II
  « une diminution importante du nombre d'exemples »
- des linguistes constatent (et déplorent) leur place décroissante
- Radermacher étudie, dans les tomes I et XVI :
  - leur provenance (sources littéraires ou autres, auteurs +/- cités)
  - la diversité des sources (nb d'exemples / nb d'œuvres ou d'auteurs)
  - leur répartition chronologique
  - la longueur (nb mots) et **le nombre d'exemples par article**
- à travers l'analyse de deux échantillons (issus du tome I et du tome XVI), elle montre :
  - que la longueur des exemples diminue effectivement
  - **que la diminution du nombre d'exemples par article est un mythe**

Le *Trésor de la langue française* : évolution au fil des tomes

Changements : les exemples (Radermacher, 2004, 2005)

- P. Imbs annonce dès le « *Au lecteur* » du tome II
  « une diminution importante du nombre d'exemples »
- des linguistes constatent (et déplorent) leur place décroissante
- Radermacher étudie, dans les tomes I et XVI :
  - leur provenance (sources littéraires ou autres, auteurs +/- cités)
  - la diversité des sources (nb d'exemples / nb d'œuvres ou d'auteurs)
  - leur répartition chronologique
  - la longueur (nb mots) et **le nombre d'exemples par article**
- à travers l'analyse de deux échantillons (issus du tome I et du tome XVI), elle montre :
  - que la longueur des exemples diminue effectivement
  - **que la diminution du nombre d'exemples par article est un mythe**

Question

Peut-on lui faire confiance ?

Échantillonnage et conclusions de Radermacher

Démarche : à partir du TLFi. . .

- reproduire l'expérience sur les échantillons de Radermacher
- reproduire automatiquement l'expérience sur l'intégralité des tomes I et XVI pour tester la « fiabilité » de l'échantillonnage et des conclusions

Reproductibilité : constitution des échantillons

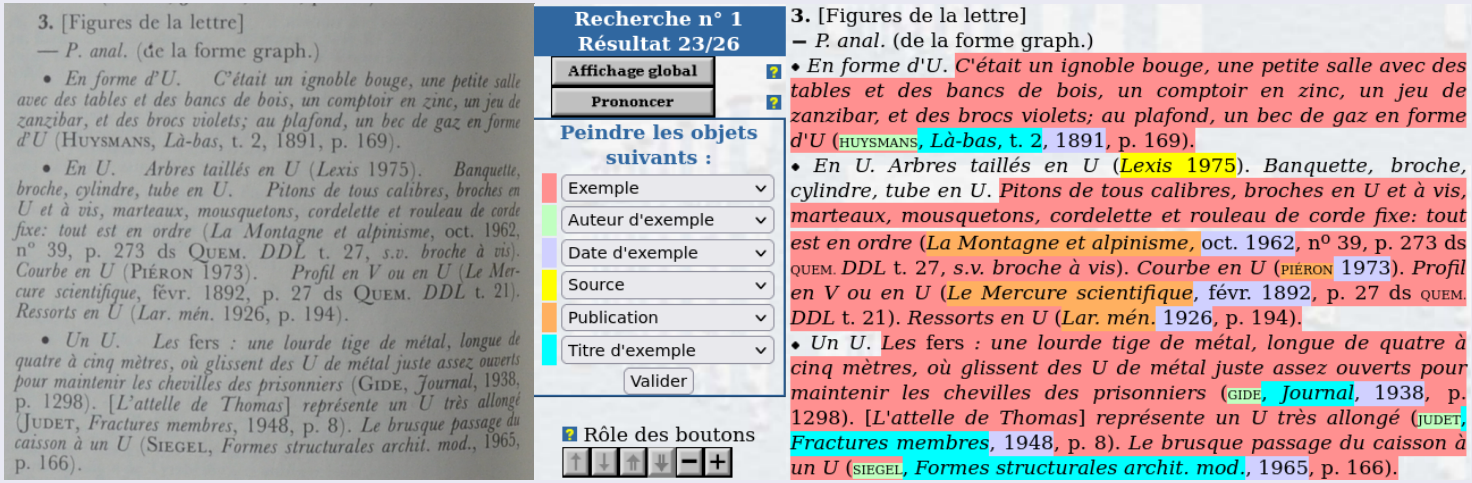
- Pour l'étude des sources :
  - tome I : « 500 exemples de la lettre A »
  - tome XVI : « 550 exemples de la lettre U »

  → zone contiguë (quel point de départ ?) ou tirage aléatoire ?
  pourquoi 2 échantillons de tailles différentes ?
- **Pour les exemples**, 2 échantillons de 100 articles :
  - T. I : « le hasard a fait le choix de » la tranche *abatture – abolir*
  - T. XVI : la tranche *U, u, lettre – unitarisme*

  → tranches contiguës, 100 exemples chacune, *préf.*, *suff.* et *élém. form./de compos.* exclus, cf. Radermacher (2004), renvois exclus (bien sûr)
  Pourquoi début de lettre (T. XVI : *U*) *vs.* tirage aléatoire de l'article initial (T. I : *abatture*) ?

Échantillonnage et conclusions de Radermacher

Reproductibilité : comptage



Nombre total d'exemples, par tranche			
Tome	**Tranche**	**Radermacher**	**FS**
T. I	*abatture – abolir*	684	639
T. XVI	*U, u – unitarisme*	795	846

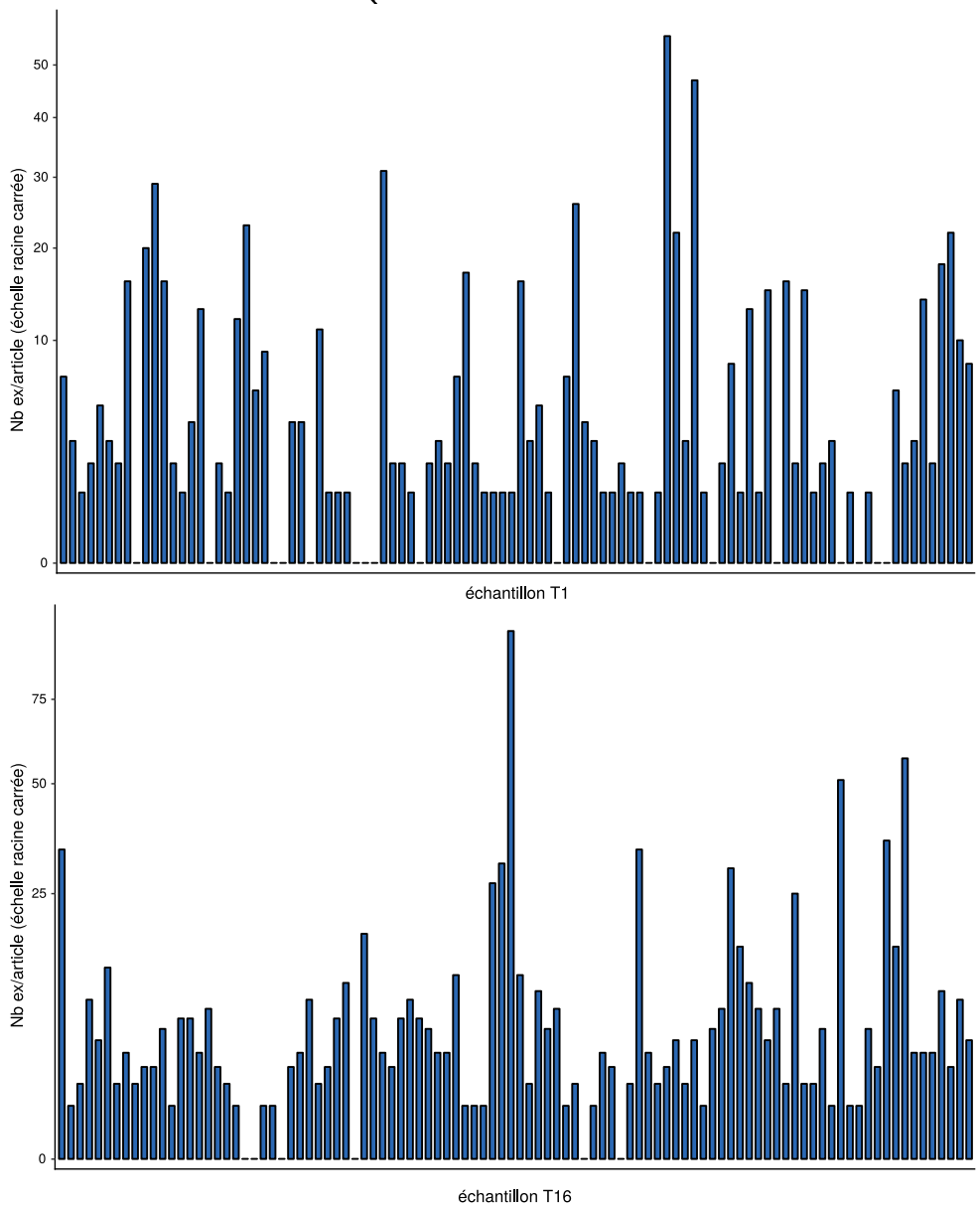
Conclusions de Radermacher

- échantillon T. I : environ 7 citations par entrée
- échantillon T. XVI : environ 8 citations par entrée

→ l'idée selon laquelle le nombre d'exemples par article aurait été la première victime des réductions effectuées à partir du tome III n'est qu'un mythe

Distributions nb ex/article T. I *vs.* T. XVI (échantillons)

Diagrammes en barre (1 barre → un article, *n* exemples)

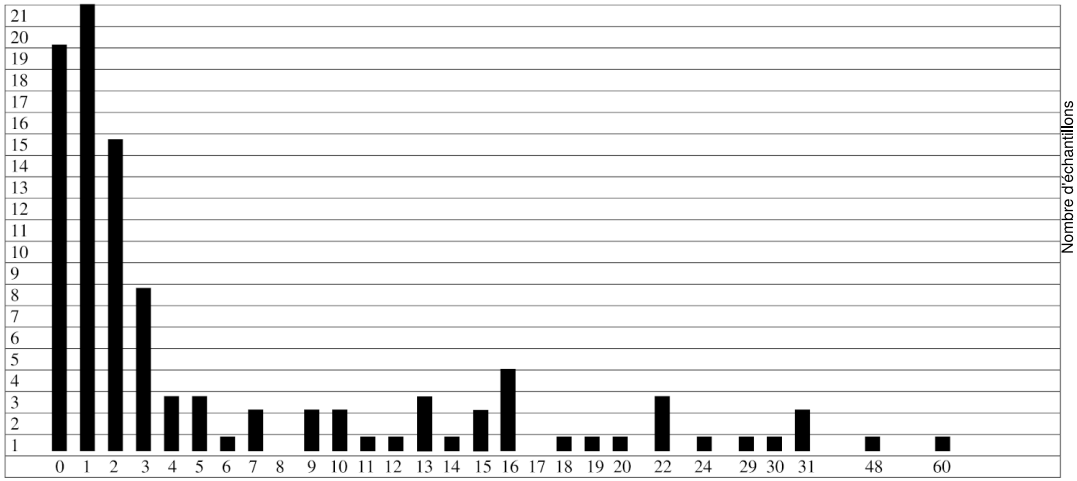


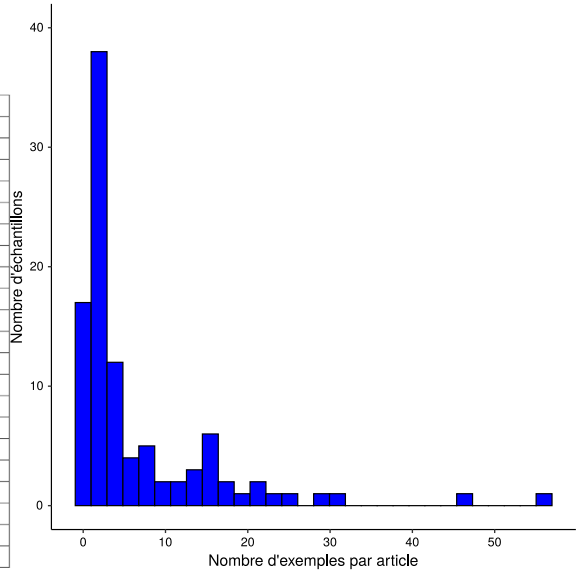
Distributions nb ex/article T. I *vs.* T. XVI (échantillons)

Radermacher

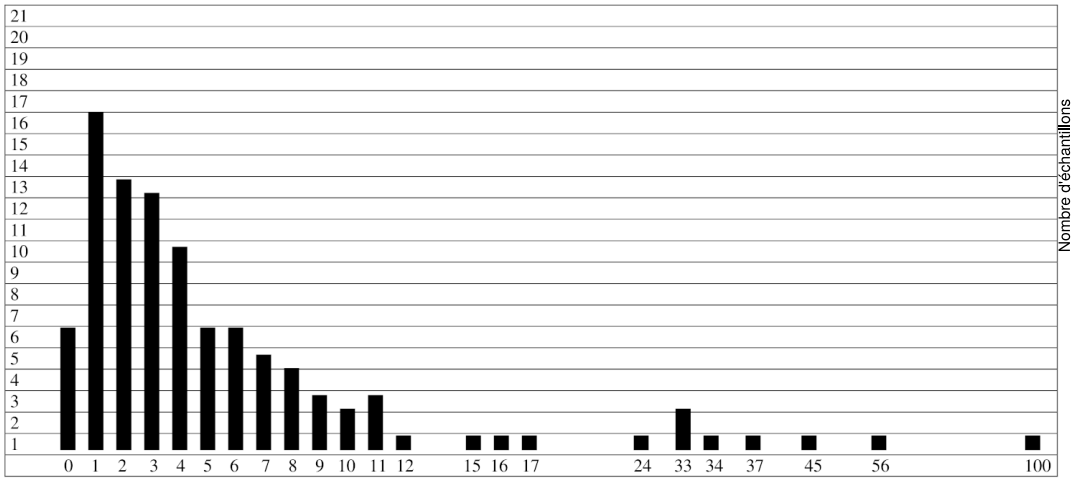
FS

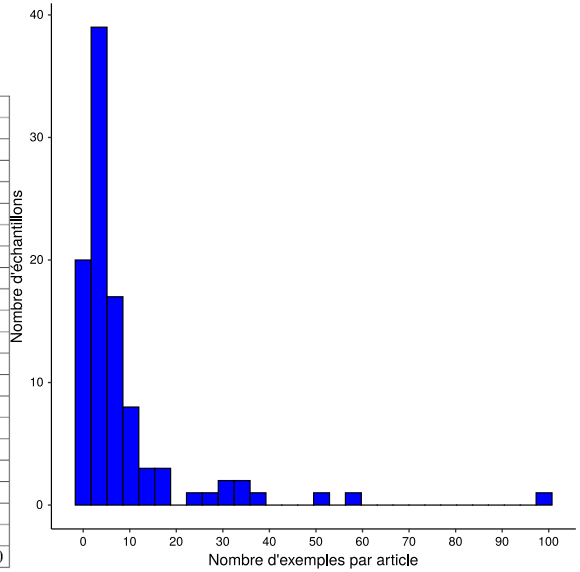
T1





T16



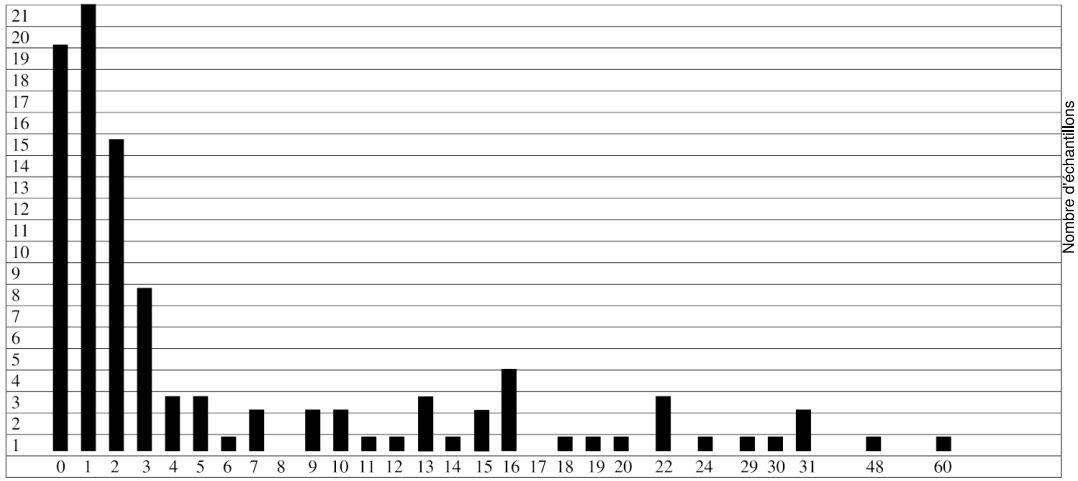


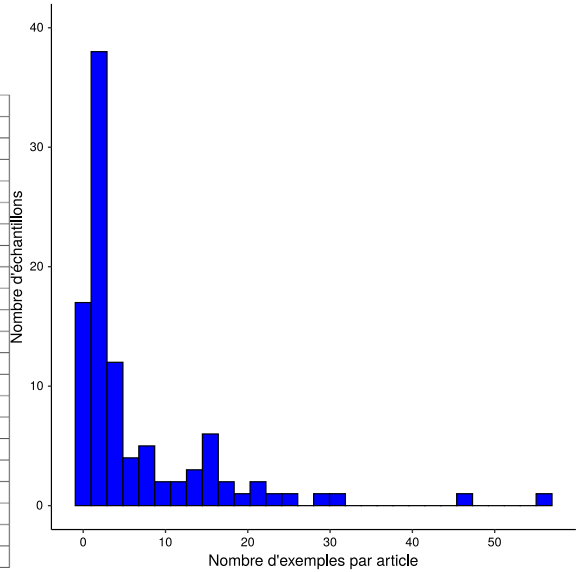
Distributions nb ex/article T. I *vs.* T. XVI (échantillons)

Radermacher

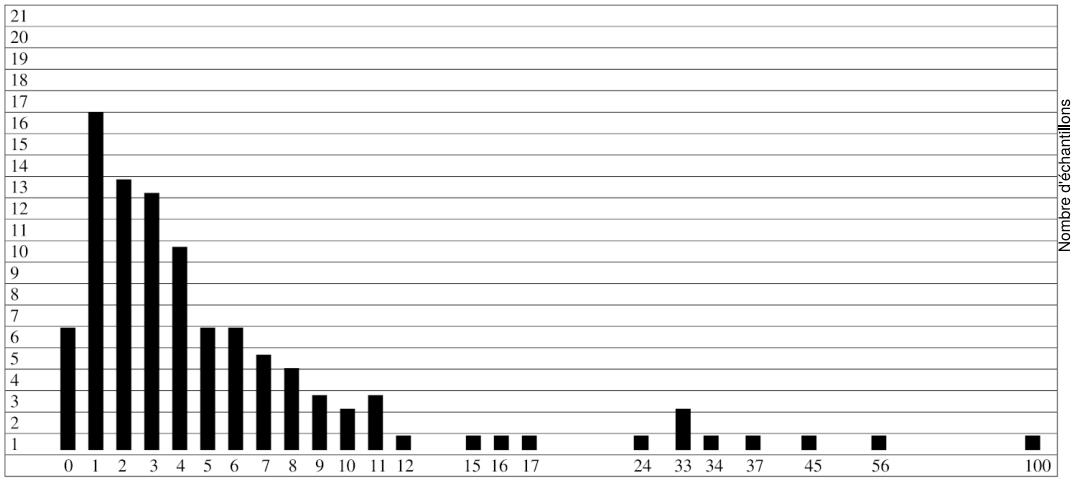
FS

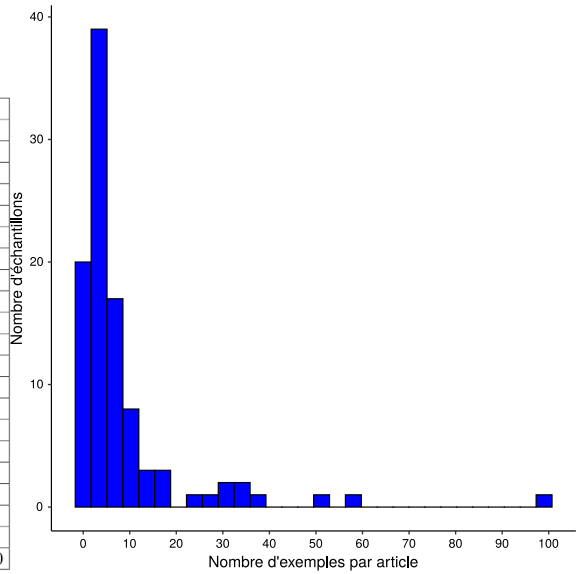
T1





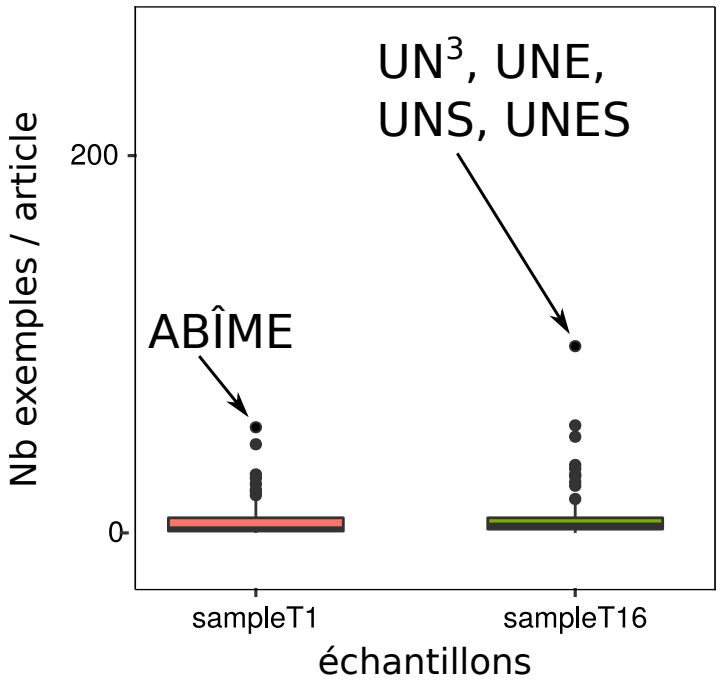
T16





distributions non normales (cf. graphiques et Shapiro-Wilk)

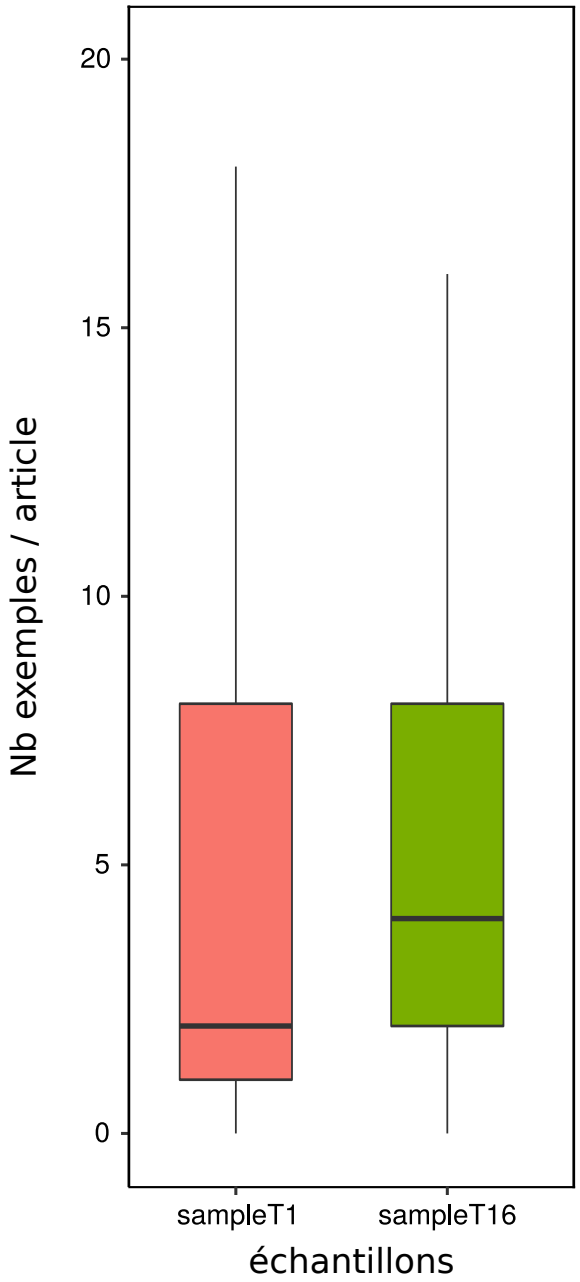
Distributions nb ex/article T. I *vs.* T. XVI (échantillons)



Tome	Min	Q1	Median	Mean	Q3	Max	σ
sampleT1	0	1	2	6.39	8	56	9.70
sampleT16	0	2	4	8.46	8	99	13.73

Pour les valeurs moyennes :
rapport sampleT16 / sampleT1 = 1,32

Distributions nb ex/article T. I *vs.* T. XVI (échantillons)



« Recadrage »
(mêmes distributions,
échelle non transformée,
valeurs extrêmes prises en compte)

Tome	Min	Q1	Median	Mean	Q3	Max	σ
sampleT1	0	1	2	6.39	8	56	9.70
sampleT16	0	2	4	8.46	8	99	13.73

Pour les valeurs moyennes :
rapport sampleT16 / sampleT1 = 1,32

Échantillonnage : quelle probabilité d'observer quoi ?

Questions

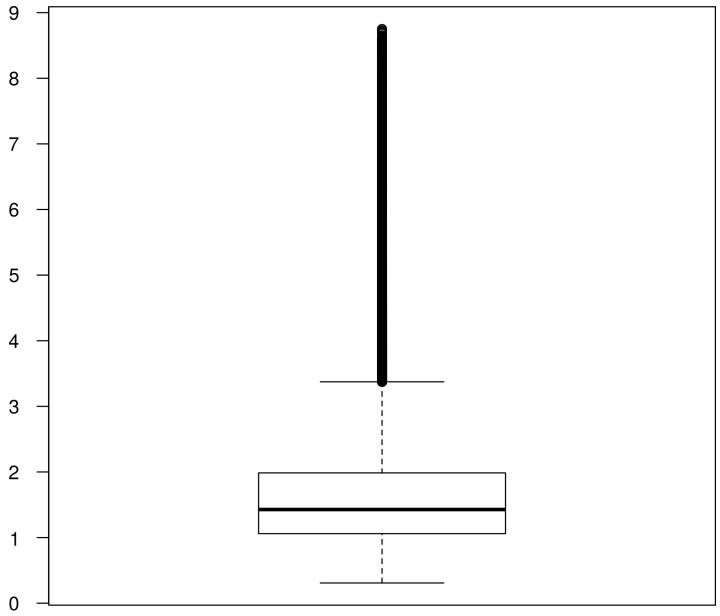
- Quelles chances a-t-on de prédire une diminution/augmentation du nombre moyen d'exemples par article T. I → T. XVI ?
- De quel ordre de grandeur ?

(paires d')échantillons

- T. I : 1 615 articles → 1 516 tranches de 100 articles
- T. XVI : 3 667 articles → 3 568 tranches de 100 articles
- 5 409 088 paires d'échantillons possibles

  Pour chaque paire :
  - comparaison des nombres moyens d'exemples par article ($<, =, >$)
  - calcul du ratio T. XVI / T. I

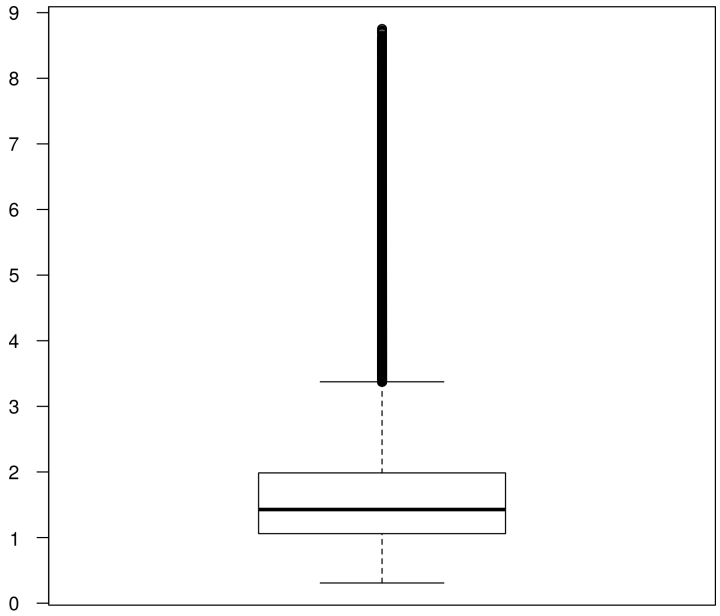
Ratio nb moyen d'exemples par article T. XVI / T. I



ratio moyenne échantillon T XVI / moyenne échantillon T I

Min	Q1	Median	Mean	Q3	Max
0.31	1.06	1.43	1.62	1.99	8.75

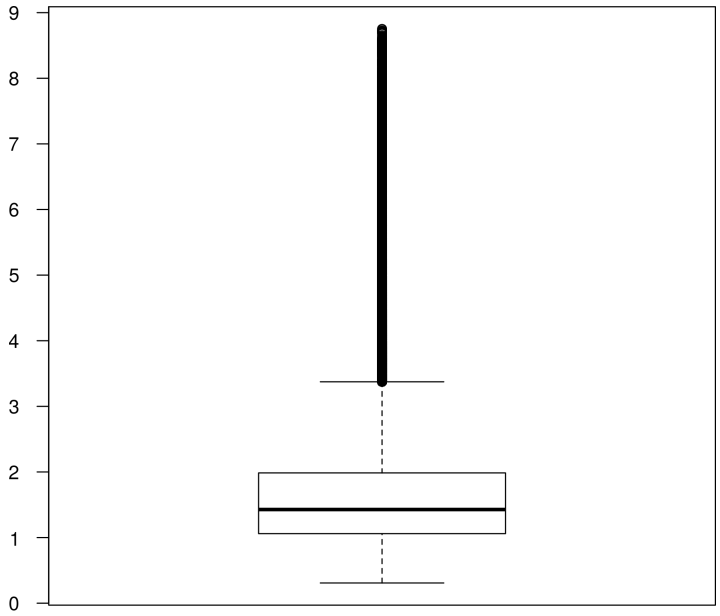
Ratio nb moyen d'exemples par article T. XVI / T. I



ratio moyenne échantillon T XVI / moyenne échantillon T I

Min	Q1	Median	Mean	Q3	Max
0.31	1.06	1.43	1.62	1.99	**8.75**

Ratio nb moyen d'exemples par article T. XVI / T. I

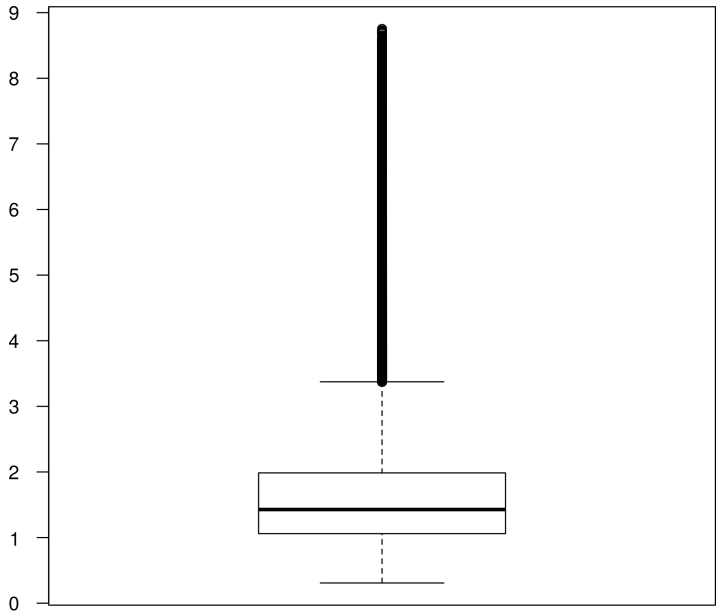


ratio moyenne échantillon T XVI / moyenne échantillon T I

Min	Q1	Median	Mean	Q3	Max
0.31	1.06	1.43	1.62	1.99	**8.75**

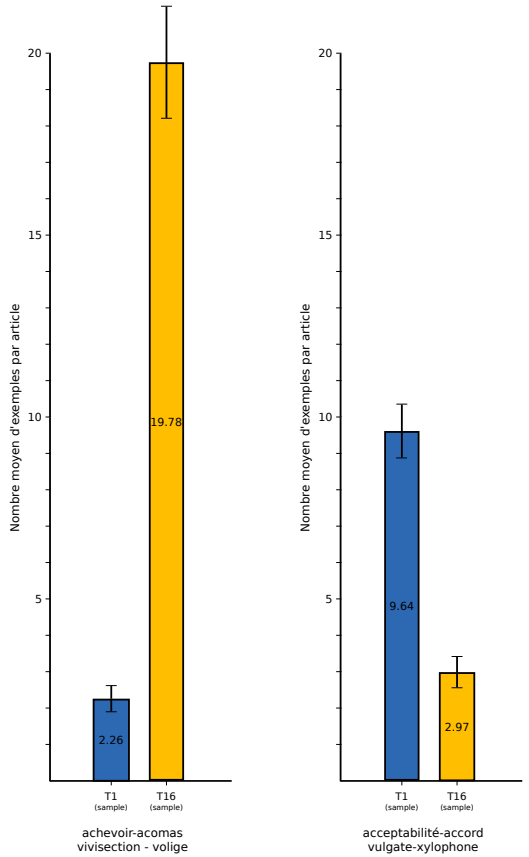
échantillon T I		échantillon T XVI		
Tranche	**Nb ex/art**	**Tranche**	**Nb ex/art**	**Ratio**
acceptabilité–accord	9.64	*vulgate–xylophone* *vulgivague–xyste*	2.97	0.31
achevoir–acomas	2.26	*vivisection–volige*	19.78	8.75

Ratio nb moyen d'exemples par article T. XVI / T. I



ratio moyenne échantillon T XVI / moyenne échantillon T I

Min	Q1	Median	Mean	Q3	Max
0.31	1.06	1.43	1.62	1.99	**8.75**



échantillon T I		échantillon T XVI		
Tranche	**Nb ex/art**	**Tranche**	**Nb ex/art**	**Ratio**
acceptabilité–accord	9.64	*vulgate–xylophone*	2.97	0.31
		vulgivague–xyste		
achevoir–acomas	2.26	*vivisection–volige*	19.78	8.75

Ratio nb moyen d'exemples par article T. XVI / T. I

Ratio entre nb moyen d'exemples par article

T XVI / TI	Nombre	%
< 1	1 147 347	21,21
$= 1$	4 638	0,09
> 1	4 257 103	78,70

Ratio nb moyen d'exemples par article T. XVI / T. I

Ratio entre nb moyen d'exemples par article

T XVI / TI	Nombre	%
< 1	1 147 347	21,21
$= 1$	4 638	0,09
> 1	4 257 103	78,70

. . . avec différence significative (Wilcoxon-Mann-Whitney)

T XVI / T I	Nombre	%
< 1	56 068	1,04
> 1	514 248	9,50

Ratio nb moyen d'exemples par article T. XVI / T. I

Ratio entre nb moyen d'exemples par article

T XVI / TI	Nombre	%
< 1	1 147 347	21,21
$= 1$	4 638	0,09
> 1	4 257 103	78,70

← *abatture–abolir vs. U, u–unitarisme*

. . . avec différence significative (Wilcoxon-Mann-Whitney)

T XVI / T I	Nombre	%
< 1	56 068	1,04
> 1	514 248	9,50

← *acceptabilité–accord vs. vulgate–xylophone* ($p < 1.26e\text{-}05$)

← *achevoir–acomas vs. vivisection–volige* ($p = .004925$)

Ratio nb moyen d'exemples par article T. XVI / T. I

Ratio entre nb moyen d'exemples par article

T XVI / TI	Nombre	%
< 1	1 147 347	21,21
$= 1$	4 638	0,09
> 1	4 257 103	78,70

← *abatture–abolir vs. U, u–unitarisme*

. . . avec différence significative (Wilcoxon-Mann-Whitney)

T XVI / T I	Nombre	%
< 1	56 068	1,04
> 1	514 248	9,50

← *acceptabilité–accord vs. vulgate–xylophone* ($p < 1.26e\text{-}05$)

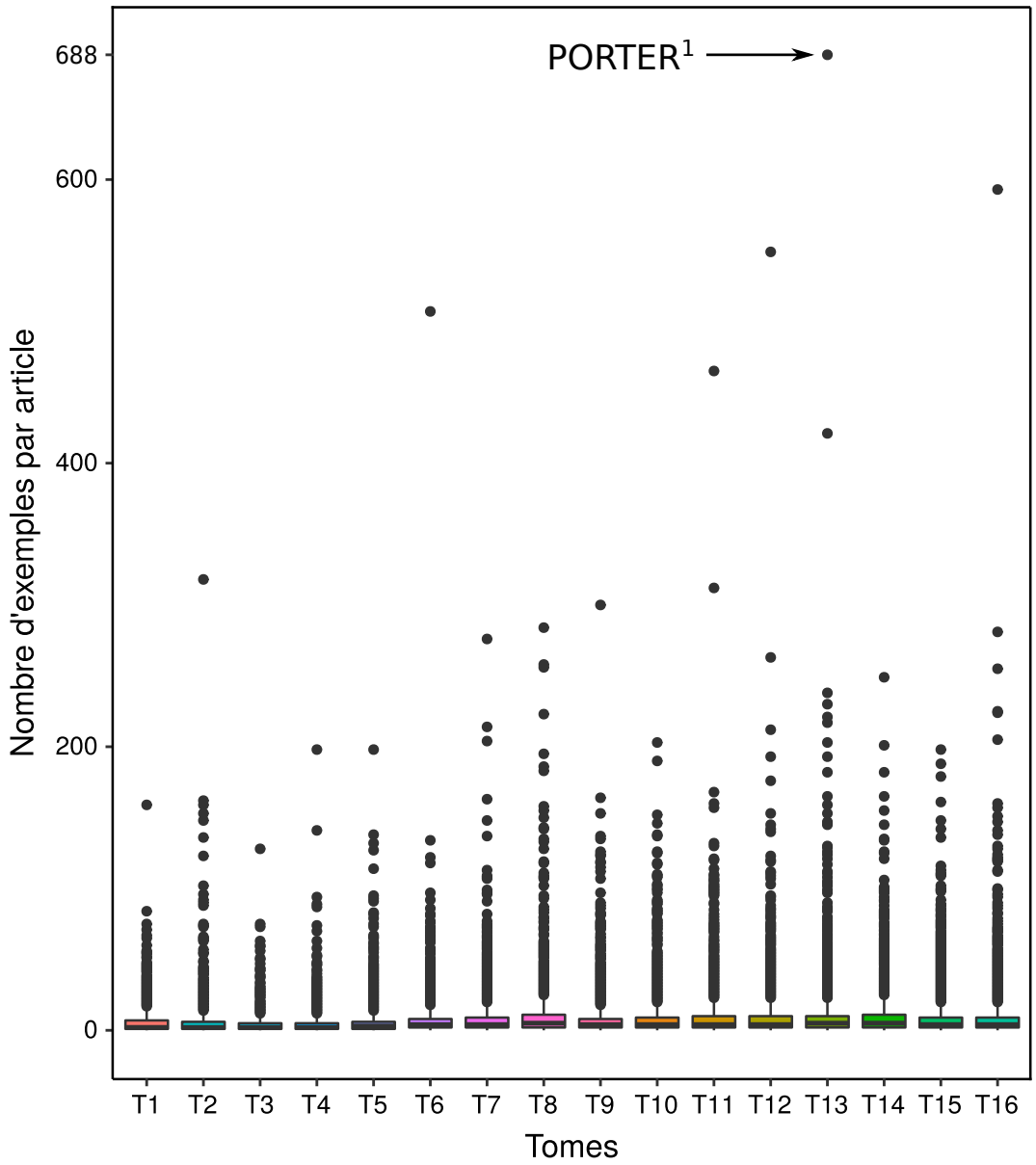
← *achevoir–acomas vs. vivisection–volige* ($p = .004925$)

Mais en vrai. . .

il y moins d'exemples par article dans T. XVI que dans T. I, ou pas ?

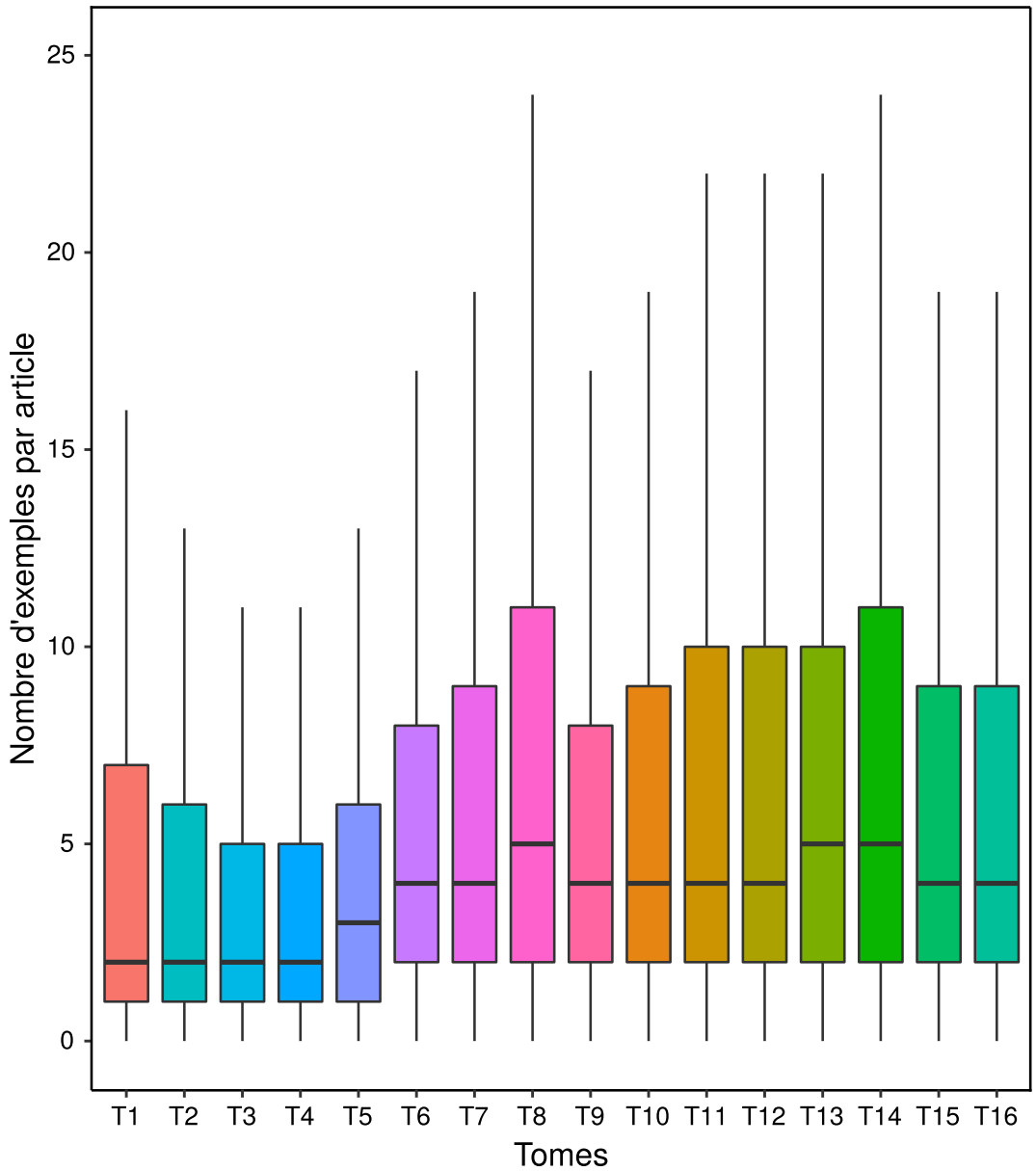
God only knows: T. I → T. XVI

Évolution du nombre d'exemples par article pour les 16 tomes

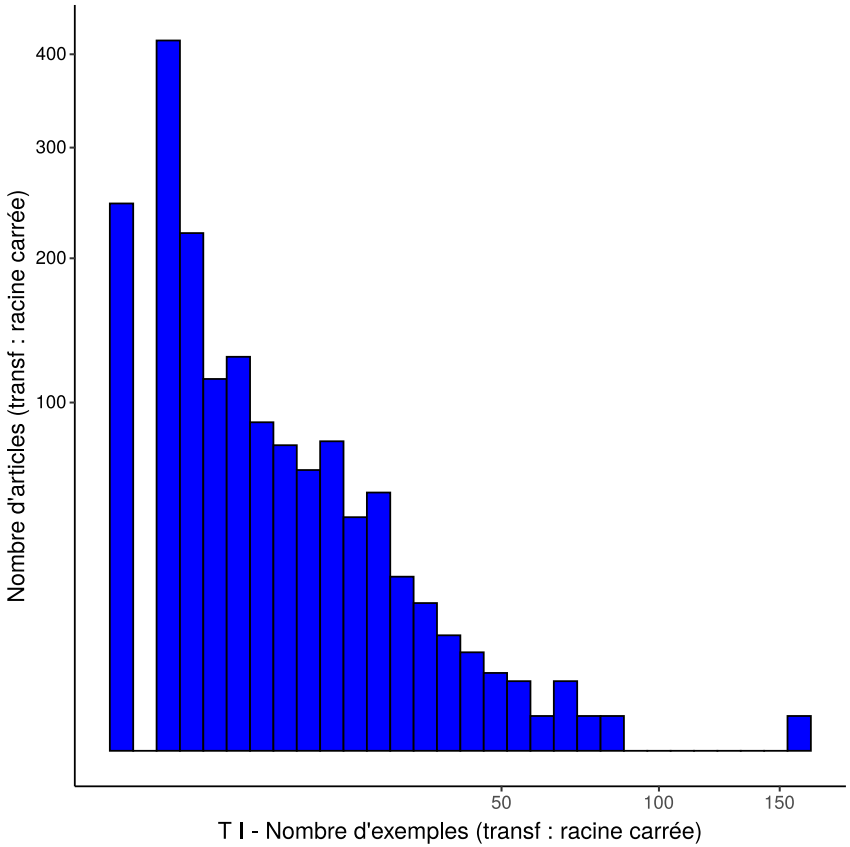


God only knows: T. I → T. XVI

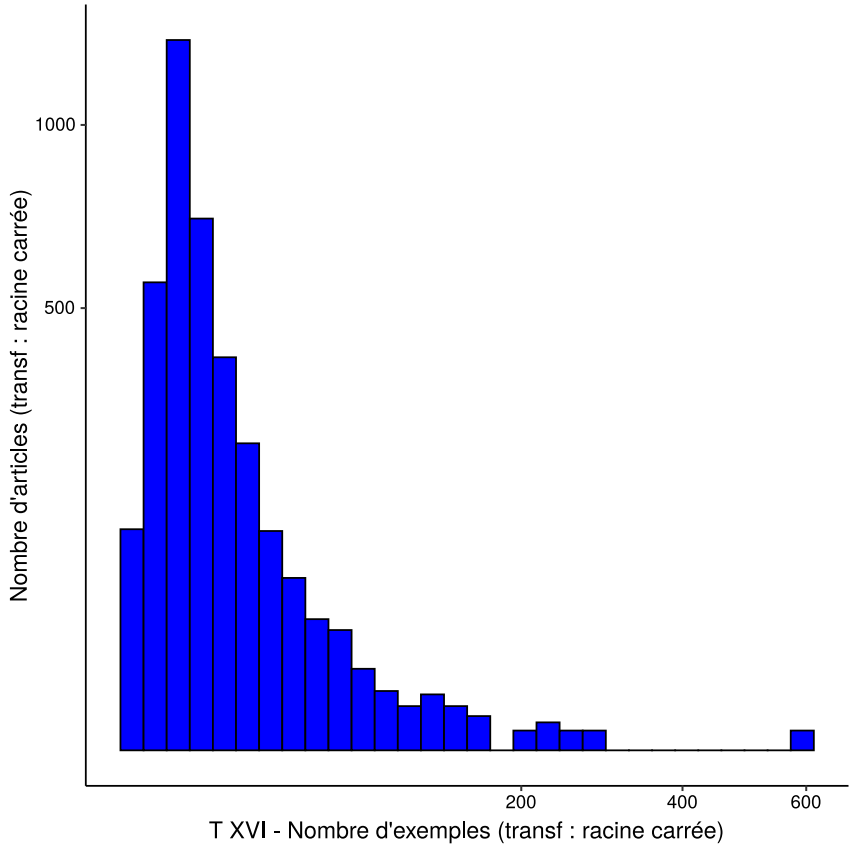
Évolution du nombre d'exemples par article pour les 16 tomes (recadrage)



God only knows: distributions tout T. I et tout T. XVI



Min	Q1	Median	Mean	Q3	Max	σ
0	1	2	6.34	7	159	10.34



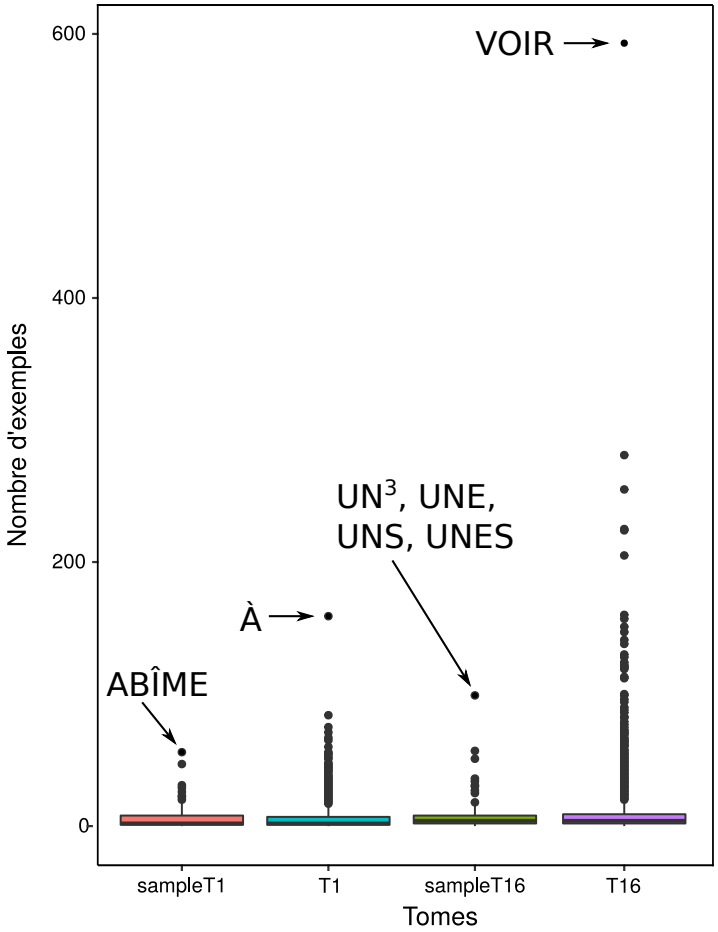
Min	Q1	Median	Mean	Q3	Max	σ
0	2	4	9.04	9	593	19.16

Distributions non normales.

Nb ex/art T. XVI > T. I (ratio moyennes : 1,43)

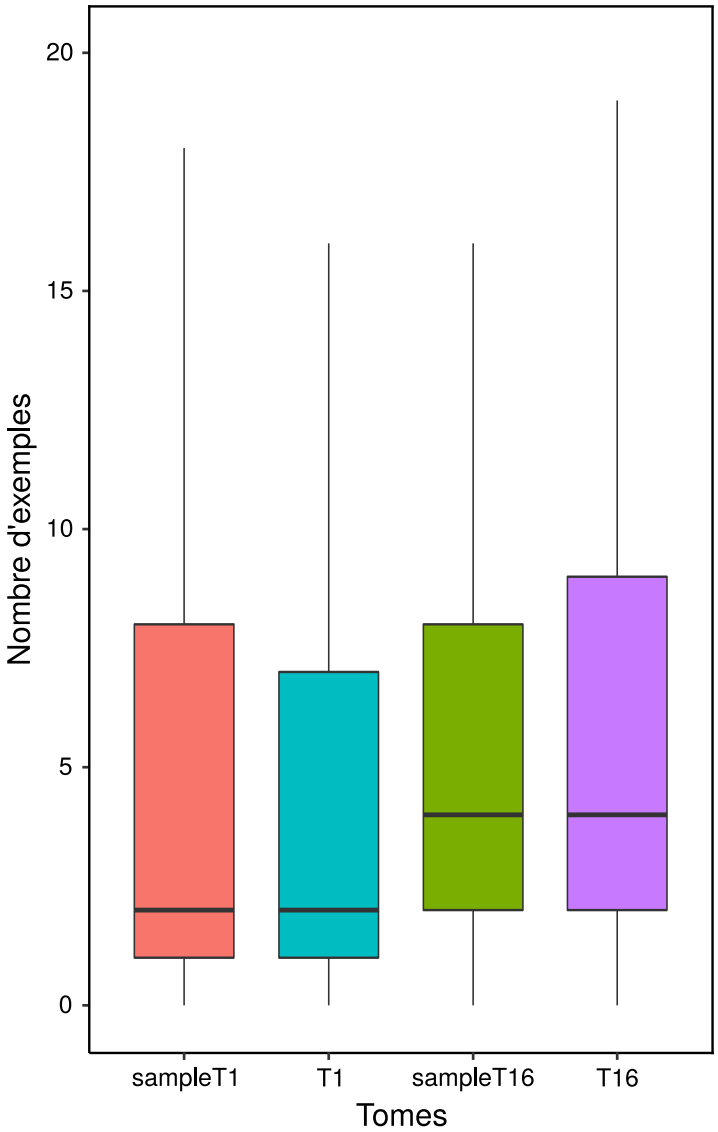
différence significative (Wilcoxon p-value < 2.2e-16)

God only knows: échantillons T. I et T. XVI *vs.* totalité



Tome	Min	Q1	Median	Mean	Q3	Max	σ
T1	0	1	2	6.34	7	159	10.34
sampleT1	0	1	2	6.39	8	56	9.70
T16	0	2	4	9.04	9	593	19.16
sampleT16	0	2	4	8.46	8	99	13.73

God only knows: échantillons T. I et T. XVI *vs.* totalité



Tome	Min	Q1	Median	Mean	Q3	Max	σ
T1	0	1	2	6.34	7	159	10.34
sampleT1	0	1	2	6.39	8	56	9.70
T16	0	2	4	9.04	9	593	19.16
sampleT16	0	2	4	8.46	8	99	13.73

Expériences de Radermacher (2004, 2005) : conclusions

Finalement. . .

- Radermacher avait raison : pas moins d'exemples par article dans T. XVI que dans T. I
- avec sa méthode, elle avait 21% de chances de conclure l'inverse
- elle n'affirme pas qu'il y a plus d'exemples dans T. XVI
  (ce qui est sage, mais dommage)
- en observant sa distribution/les articles, elle fournissait une explication valide : peu d'articles avec énormément d'exemples dans T. XVI, mais beaucoup moins d'articles sans exemple que dans T. I

Expériences de Radermacher (2004, 2005) : conclusions

Finalement. . .

- Radermacher avait raison : pas moins d'exemples par article dans T. XVI que dans T. I
- avec sa méthode, elle avait 21% de chances de conclure l'inverse
- elle n'affirme pas qu'il y a plus d'exemples dans T. XVI
  (ce qui est sage, mais dommage)
- en observant sa distribution/les articles, elle fournissait une explication valide : peu d'articles avec énormément d'exemples dans T. XVI, mais beaucoup moins d'articles sans exemple que dans T. I
- selon les échantillons, possibilité d'estimer qu'il y a :
  - 3 fois moins d'exemples dans T. XVI que dans T. I
  - (presque) 9 fois plus d'exemples dans T. XVI que dans T. I

Expériences de Radermacher (2004, 2005) : conclusions

Finalement. . .

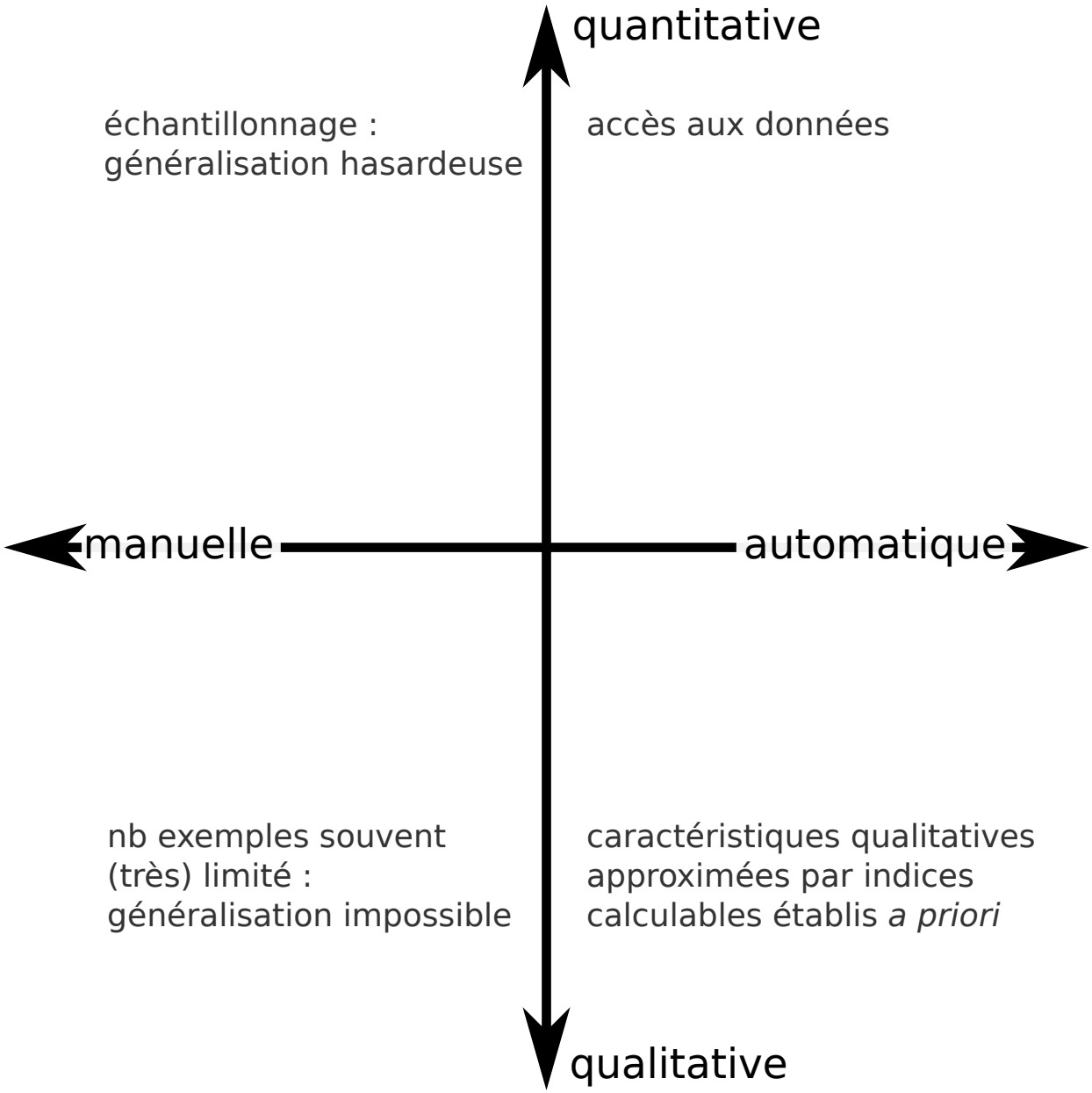
- Radermacher avait raison : pas moins d'exemples par article dans T. XVI que dans T. I
- avec sa méthode, elle avait 21% de chances de conclure l'inverse
- elle n'affirme pas qu'il y a plus d'exemples dans T. XVI
  (ce qui est sage, mais dommage)
- en observant sa distribution/les articles, elle fournissait une explication valide : peu d'articles avec énormément d'exemples dans T. XVI, mais beaucoup moins d'articles sans exemple que dans T. I
- selon les échantillons, possibilité d'estimer qu'il y a :
  - 3 fois moins d'exemples dans T. XVI que dans T. I
  - (presque) 9 fois plus d'exemples dans T. XVI que dans T. I
- en testant la significativité des différences :
  - seulement 1% de chances d'estimer que nb ex T. XVI $<$ nb ex. T. I
  - 9,5% (seulement) de montrer la supériorité du nb d'ex dans T. XVI

Conclusions : échantillonnage

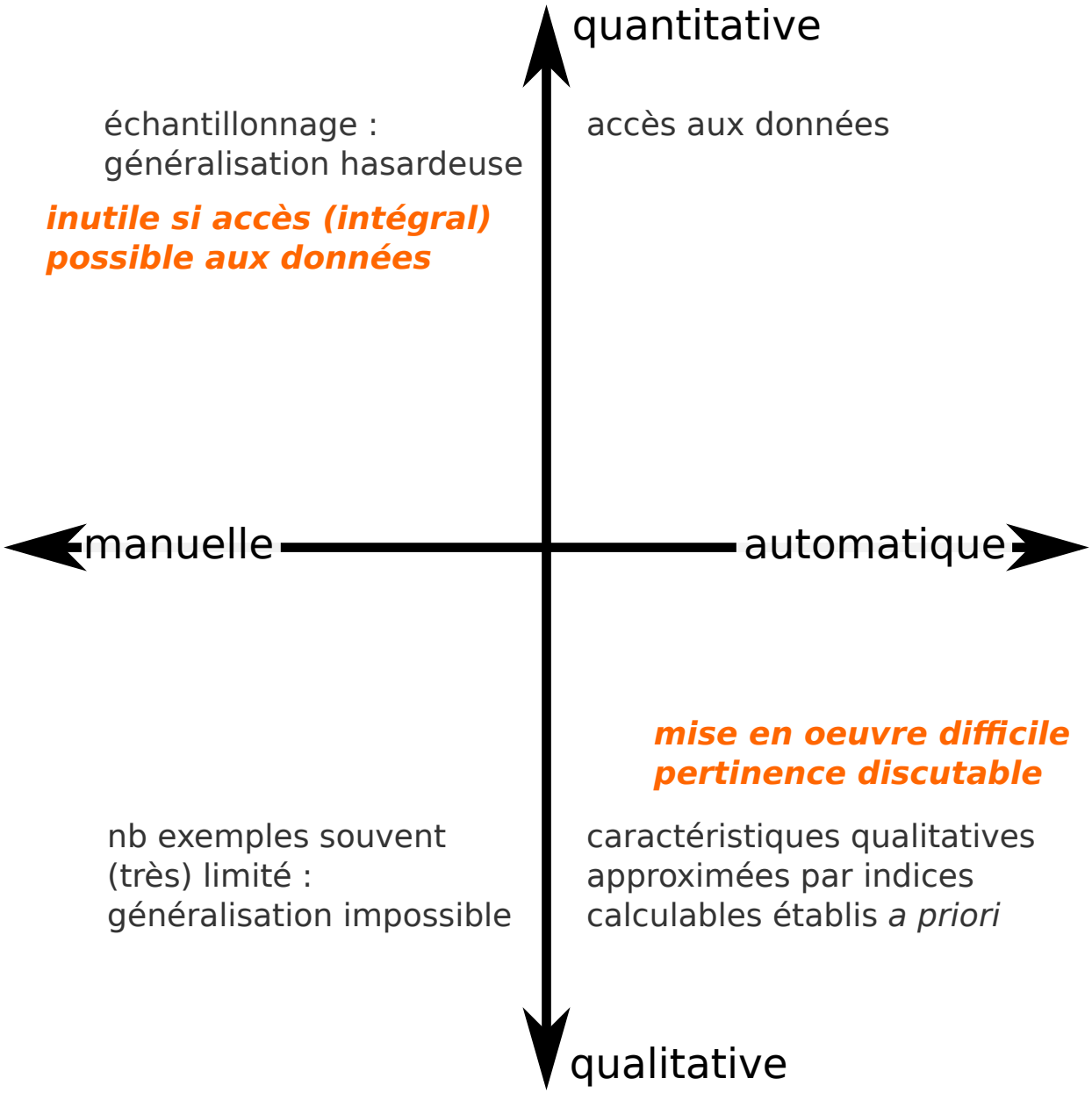
Échantillonnage

1. échantillonnage probabiliste préférable à échantillonnage par zone contiguë
2. échantillonnage sur tout le dictionnaire préférable à échantillonnage dans une tranche donnée
3. aucune garantie satisfaisante de la représentativité d'un échantillon
4. études quantitatives sur l'intégralité du dictionnaire à privilégier ! (*i.e.* pas d'échantillonnage du tout)
5. $\rightarrow$ mise en œuvre automatique (moyennant d'éventuels problèmes de droits)
6. mieux : coupler études quantitatives automatiques et études qualitatives manuelles

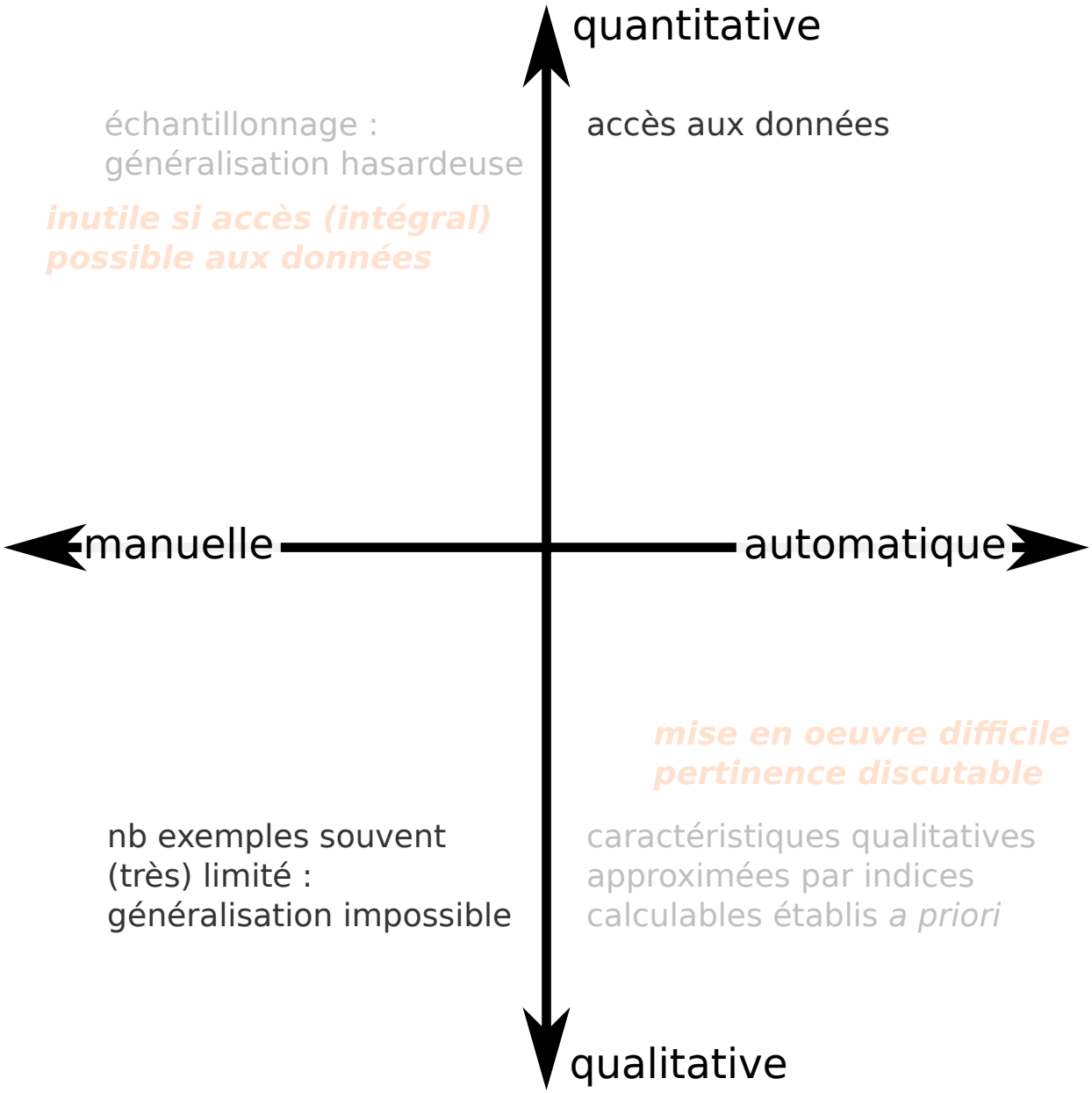
Conclusions : méthodes d'analyse



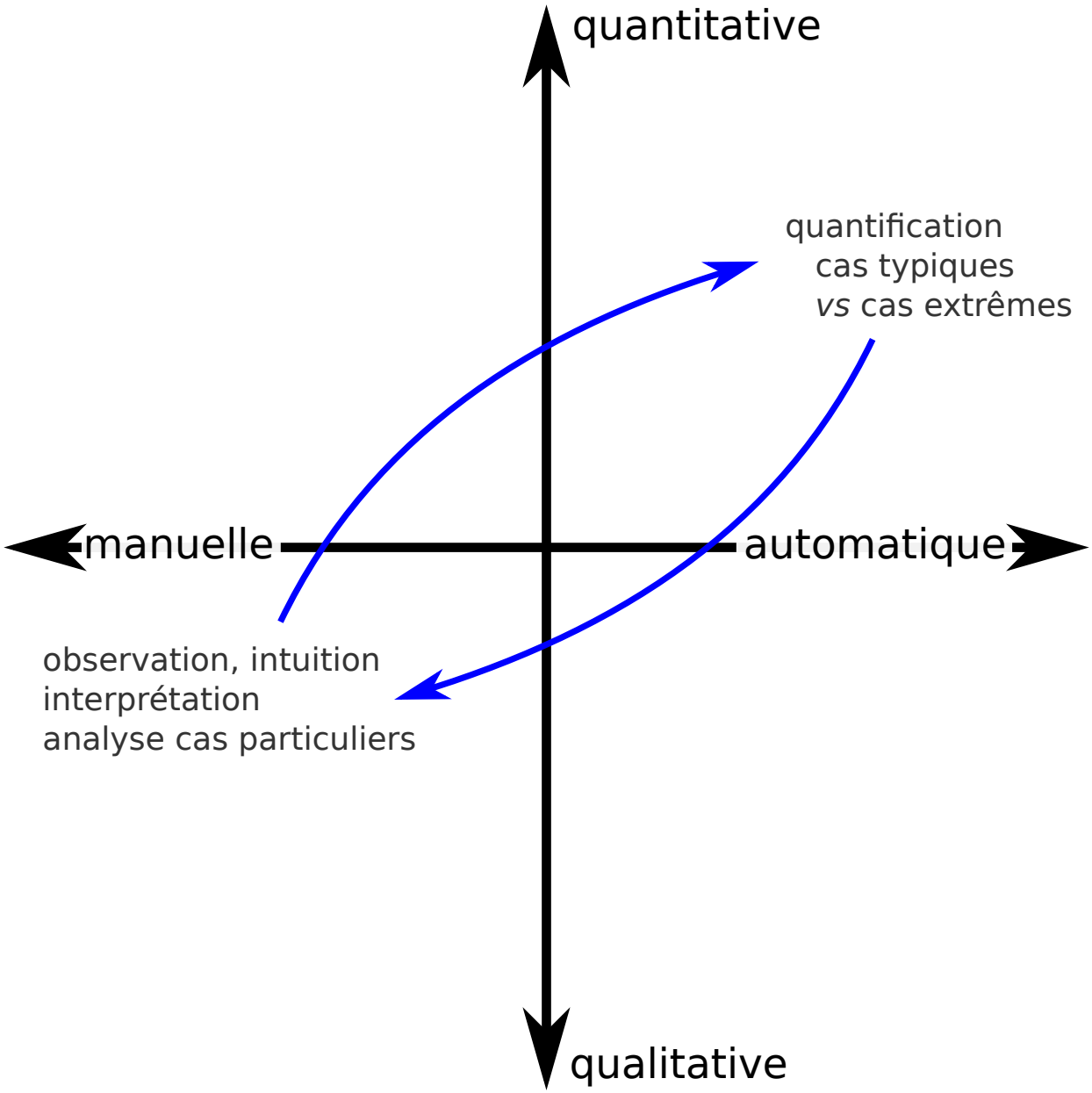
Conclusions : méthodes d'analyse



Conclusions : méthodes d'analyse

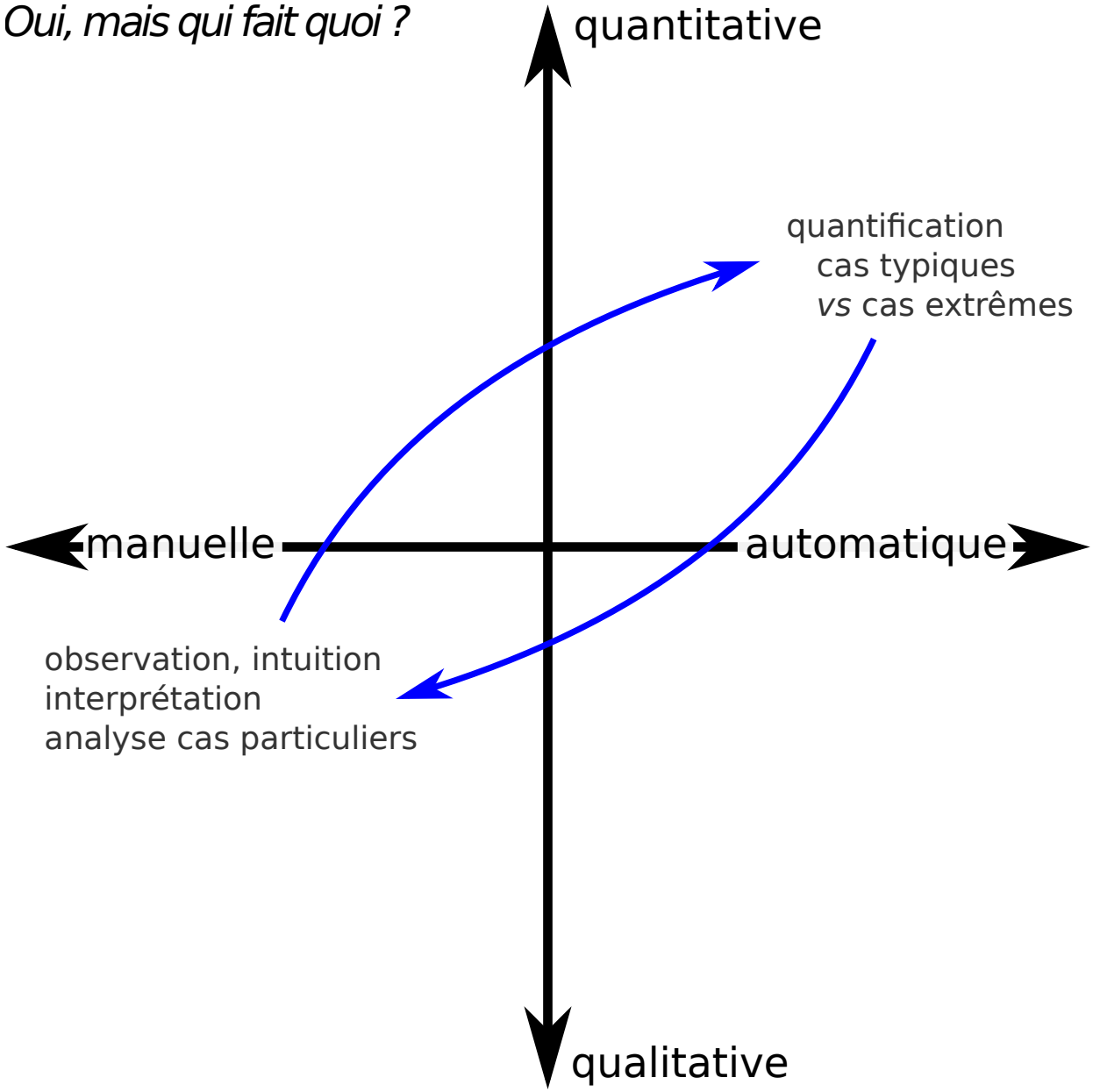


Conclusions : méthodes d'analyse



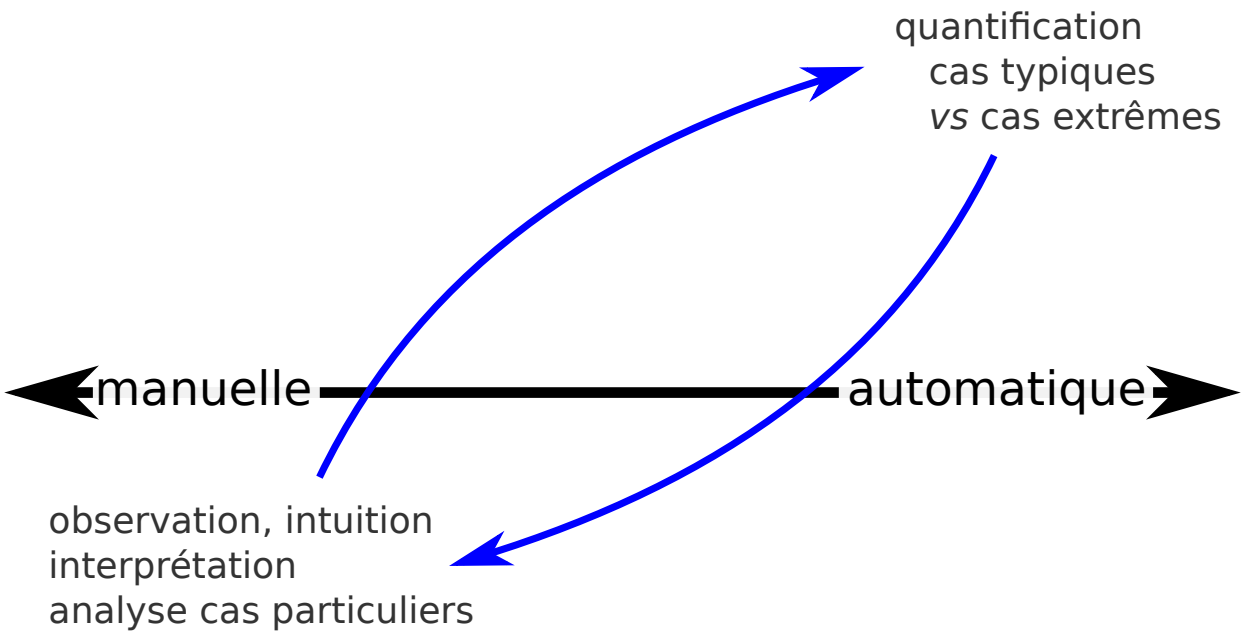
Conclusions : méthodes d'analyse

Oui, mais qui fait quoi ?



Conclusions : méthodes d'analyse

Oui, mais qui fait quoi ?



Conclusions : méthodes d'analyse

Oui, mais qui fait quoi ?

quantification
cas typiques
vs cas extrêmes

manuelle ⟷ automatique

observation, intuition
interprétation
analyse cas particuliers

Conclusions : méthodes d'analyse

Oui, mais qui fait quoi ?

quantification
cas typiques
vs cas extrêmes

manuelle ⟷ automatique

observation, intuition
interprétation
analyse cas particuliers

collaborations souhaitables (en attendant une relève polyvalente...)

Références I

Béjoint, H. (2000). *Modern Lexicography: An Introduction*. Oxford: Oxford University Press.

Berg, D., Gönnet, G., & Tompa, F. (1988). The New Oxford English Dictionary Project at the University of Waterloo. *Technical Report OED-88-01*, Centre for the New Oxford English Dictionary, University of Waterloo.

Bukowska, A. A. (2010). Sampling techniques in metalexicographic research. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress* (pp. 1258–1269). Leeuwarden/Ljouwert, The Netherlands.

Coleman, J. & Ogilvie, S. (2009). Forensic Dictionary Analysis: Principles and Practice. *International Journal of Lexicography*, 22(1), 1–22.

Références II

Corbin, P. (1990). Le monde étrange des dictionnaires (7). Logique linguistique et logique botanique : problèmes posés par la définition d'une classe de mots dérivés français. *Cahiers de lexicologie*, 57-59, 75–108.

Daxenberger, J. & Gurevych, I. (2012). A Corpus-Based Study of Edit Categories in Featured and Non-Featured Wikipedia Articles. In *Proceedings of COLING 2012* (pp. 711–726). Mumbai, India.

de Schryver, G.-M. (2005). Concurrent Over- and Under-treatment in Dictionaries – The Woordeboek van die Afrikaanse Taal as a Case in Point. *International Journal of Lexicography*, 18(1), 47–75.

de Schryver, G.-M. (2022). Metalexicography: an existential crisis. In *Proceedings of the 20th EURALEX International Congress* (pp. 196–206). Mannheim, Germany.

Références III

Elchacar, M. (2019). Comparaison du traitement lexicographique des appellations des identités de genre non traditionnelles dans les dictionnaires professionnels et profanes. *Études de linguistique appliquée*, 194(2), 177–191.

Freeman, H. (1963). *Introduction to statistical inference*. Rearing, MA: Addison-Wesley Publishing Company.

Gao, Y. (2012). Online English Dictionaries: Friend or Foe. In *Proceedings of the 15th EURALEX International Congress* (pp. 422–433). Oslo, Norway.

Hanks, P. (2012). Corpus evidence and electronic lexicography. In S. Granger & M. Paquot (Eds.), *Electronic Lexicography* (pp. 57–82). Oxford: Oxford University Press.

Hartmann, R. R. K. (2001). *Teaching and Researching Lexicography*. London: Routledge.

Références IV

Josselin-Leray, A. (2010). Affiner la description des termes dans les dictionnaires généraux : l'apport d'un corpus de vulgarisation. *Lexis*, 4, 65–104.

Kilgarriff, A. (2005). If dictionaries are free, who will buy them? *Kernerman Dictionary News*, 13, 17–19.

Kittur, A. & Kraut, R. E. (2008). Harnessing the Wisdom of Crowds in Wikipedia: Quality through Coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08 (pp. 37–46). New York, NY, USA: Association for Computing Machinery.

Lew, R. (2014). User-generated content (UGC) in online English dictionaries. *OPAL*, 4, 8–26.

Martinez, C. (2013). La comparaison de dictionnaires comme méthode d'investigation lexicographique. *Lexique*, 21, 193–220.

Références V

Meyer, C. M. (2013). *Wiktionary: The Metalexicographic and the Natural Language Processing Perspective*. PhD thesis, Technische Universität Darmstadt.

Nagao, M., Tsujii, J., Ueda, Y., & Takiyama, M. (1980). An attempt to computerized dictionary data bases. In *Proceedings of COLING 1980* (pp. 534–542). Tokyo, Japan.

Nesi, H. (2008). Dictionaries in electronic form. In A. P. Cowie (Ed.), *The Oxford History of English Lexicography* (pp. 458–478). Oxford: Oxford University Press.

Osselton, N. E. (2007). Alphabet Fatigue and Compiling Consistency in Early English Dictionaries. In J. Considine & G. Iamartino (Eds.), *Words and Dictionaries from the British Isles in Historical Perspective* (pp. 81–90). Newcastle: Cambridge Scholars Publishing.

Références VI

Radermacher, R. (2004). *Le Trésor de la Langue Française. Une étude historique et lexicographique*. PhD thesis, Université Marc Bloch, Strasbourg.

Radermacher, R. (2005). Les citations dans le *Trésor de la langue française*. In M. Heinz (Ed.), *L'exemple lexicographique dans les dictionnaires français contemporains*, volume 128 of *Lexicographica Series Maior* (pp. 215–229). Berlin, Boston: De Gruyter.

Rundell, M. (2014). Macmillan English Dictionary: The End of Print? *Slovenščina 2.0*, 2(2), 1–14.

Rundell, M. (2017). Dictionaries and crowdsourcing, wikis, and user-generated content. In P. Hanks & G.-M. de Schryver (Eds.), *International Handbook of Modern Lexis and Lexicography*. Berlin, Heidelberg: Springer.

Références VII

Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: Where will it all end? In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A Taste for Corpora. In honour of Sylviane Granger* (pp. 257–282). John Benjamins.

Rundell, M. & Stock, P. (1992). The corpus revolution. *English Today*, 30, 9–14.

Sajous, F. (2023). Quantité et qualité dans le Wiktionnaire : de la diversité... à la rigueur ? *Linx*, 86.

Sajous, F. & Humbley, J. (2022). Mesures d'isolement sanitaire dans Wiktionnaire et Wikipédia : néologie et lexicographie ou néonymie et terminographie ? *Estudios Románicos*, 31, 175–201.

Sajous, F. & Josselin-Leray, A. (2022). Issues in Collaborative and Crowdsourced Lexicography. In H. Jackson (Ed.), *The Bloomsbury Handbook of Lexicography* (pp. 343–358). London: Bloomsbury Publishing.

Références VIII

Sajous, F., Josselin-Leray, A., & Hathout, N. (2018). Définir la néologie terminologique dans les dictionnaires généraux : le domaine de l'informatique analysé par « les foules » et par les professionnels... de la lexicographie. In *4ème Congrès international de néologie des langues romanes (CINEO 2018)* Lyon, France.

Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2005). Assessing information quality of a community-based encyclopedia. In *Proceedings of the 2005 International Conference on Information Quality (ICIQ 2005)* (pp. 442–454). Cambridge, MA.

Trap-Jensen, L. (2018). Lexicography between NLP and Linguistics: Aspects of Theory and Practice. In *Proceedings of the 18th EURALEX International Congress* (pp. 25–37). Ljubljana.

Vincent, N. (2022). Faut-il adapter les dictionnaires à l'air du temps ? Proposition d'un traitement polyphonique du mot *woke*. Regards linguistiques sur des mots polémiques, *Circula*, 15, 122–145.

Références IX

Wilkinson, D. M. & Huberman, B. A. (2007). Assessing the value of cooperation in Wikipedia. *First Monday*, 12(4).

Wolfer, S. & Müller-Spitzer, C. (2016). How Many People Constitute a Crowd and What Do They Do? Quantitative Analyses of Revisions in the English and German Wiktionary Editions. *Lexicos*, 26.