Détection des innovations lexicales liées aux sujets d'actualité par analyse des révisions de Wiktionary

Application aux néologismes du Covid-19

Thématiques actuelles de la recherche en TAL, 22/11/2021

Franck Sajous

CLLE, CNRS & Université de Toulouse 2

Mise à jour des dictionnaires

What, When and How? – the Art of Updating an Online Dictionary

"Ideally, all entries in a dictionary need to be revised from time to time, or at least to be checked to see if revisions are needed. And any part of the entry may be subject to change." (Lorentzen and Trap-Jensen, 2016)

Mise à jour des dictionnaires

What, When and How? – the Art of Updating an Online Dictionary

"Ideally, all entries in a dictionary need to be revised from time to time, or at least to be checked to see if revisions are needed. And any part of the entry may be subject to change." (Lorentzen and Trap-Jensen, 2016)

Interrogatifs manquants : *who?* et *how much?*

Mise à jour des dictionnaires

What, When and How? – the Art of Updating an Online Dictionary

"Ideally, all entries in a dictionary need to be revised from time to time, or at least to be checked to see if revisions are needed. And any part of the entry may be subject to change." (Lorentzen and Trap-Jensen, 2016)

Interrogatifs manquants : *who?* et *how much?*

What: description de la langue *vs.* description du monde

- unités linguistiques : néologismes, changements orthographiques, etc.
- discours métalinguistiques (*e.g.* marquage diachronique, diastratique, etc.)
- concepts (référents) : changements sociétaux (nom des diplômes, des monnaies, des pays, déf. de *mariage*, etc.) et technologiques, évolutions des connaissances, etc.
- « normes » du moment : rémanence de stéréotypes racistes et sexistes (Farina, 2005), morale religieuse : *"genre de libertinage solitaire nuisible à la santé"*, *"péché/vice contre nature"* dans *Littré* et *DAF8* mais réprobation observable encore récemment (Boisson, 2000), féminisation...

Mise à jour des dictionnaires : *what* (suite)

Nouvelle entrée = néologisme ?

Inclusion de néologismes dans le dictionnaire : une évidence ?

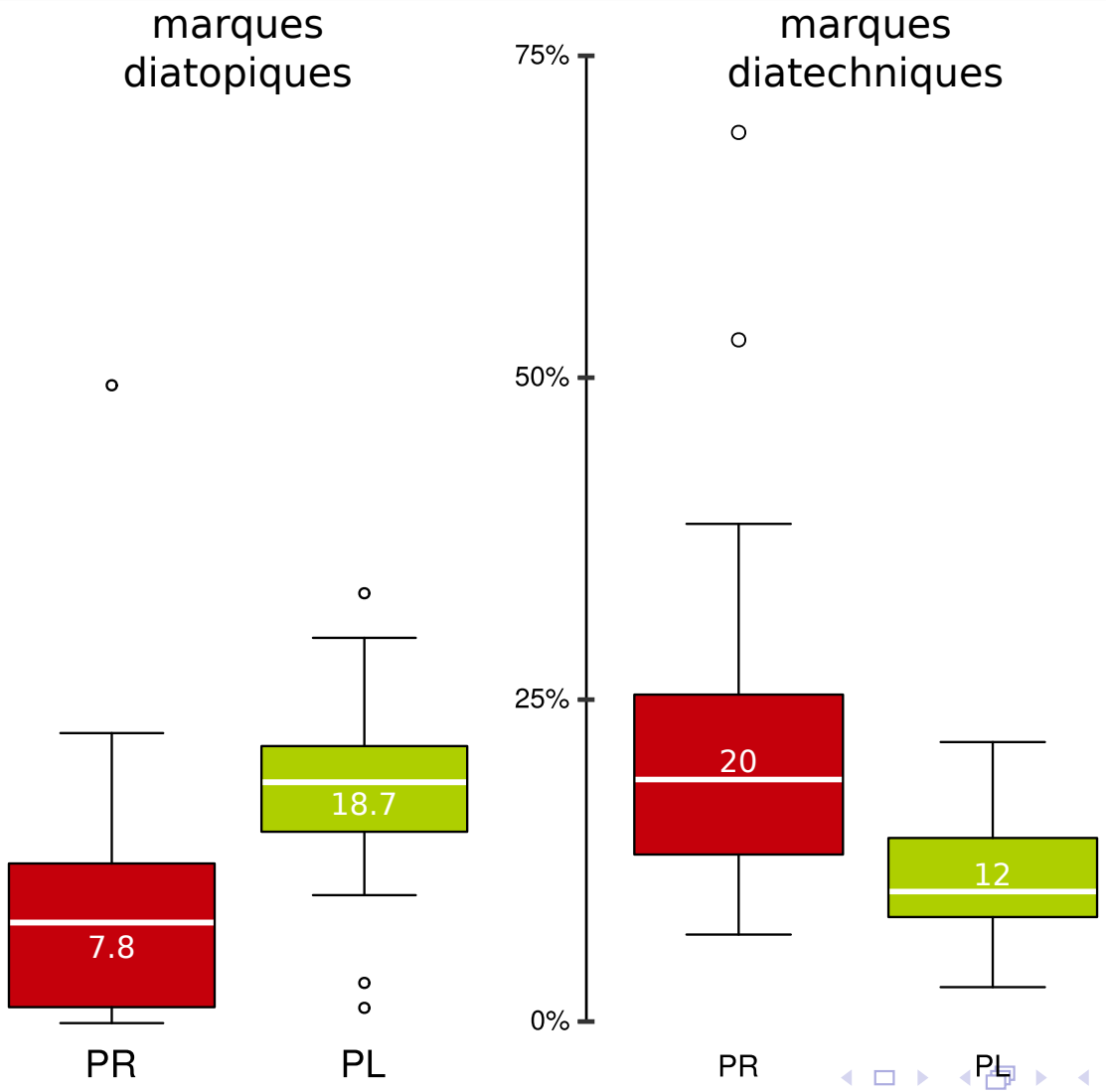
"new entries and new features attract more attention from users and sponsors alike, because novelties are generally considered more interesting than maintenance, updating is often given less priority and tends to be neglected" (Lorentzen and Trap-Jensen, 2016)

ou un paradoxe ?

- critères d'inclusion : fréquence et installation dans l'usage (entre autres)
- entrée dans un dictionnaire → achèvement du processus de lexicalisation (Mortureux, 2011)
- dictionnaire = corpus d'exclusion pour la recherche de néologismes (Sablayrolles, 2008)
- inclusion de *buzzwords*, *e.g. iel* (qui **n'est pas** entré dans le PR, mais dans le *Dico en ligne* des éditions *Le Robert*)

Mise à jour des dictionnaires : *what* (suite)

% variantes diatopiques / termes parmi les nouvelles entrées (Sajous and Martinez, 2021)

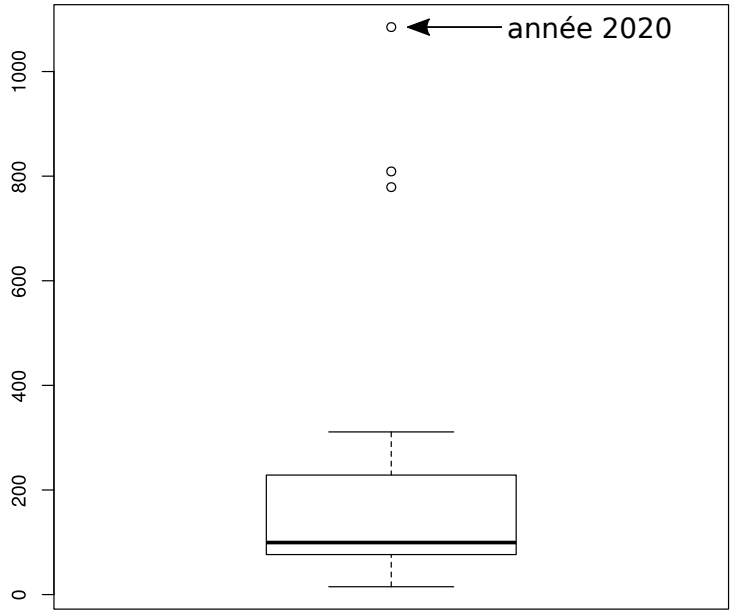


Mise à jour des dictionnaires : *when (& how much)?*

Fréquence des mises à jour

- *DAF8* : 1932-1935 → *DAF9* : en cours de rédaction
- mises à jour annuelles, *e.g. Petit Larousse* et *Petit Robert*
- *OED* : mises à jour trimestrielles + mises à jour exceptionnelles

  `https://public.oed.com/updates/`
- mises à jour continue (dictionnaires en ligne), *e.g. Usito*



***OED* : nouveaux sens**	
Min.	15.00
1st Qu.	76.75
Median	99.50
Mean	228.50
3rd Qu.	213.25
Max.	1085.00

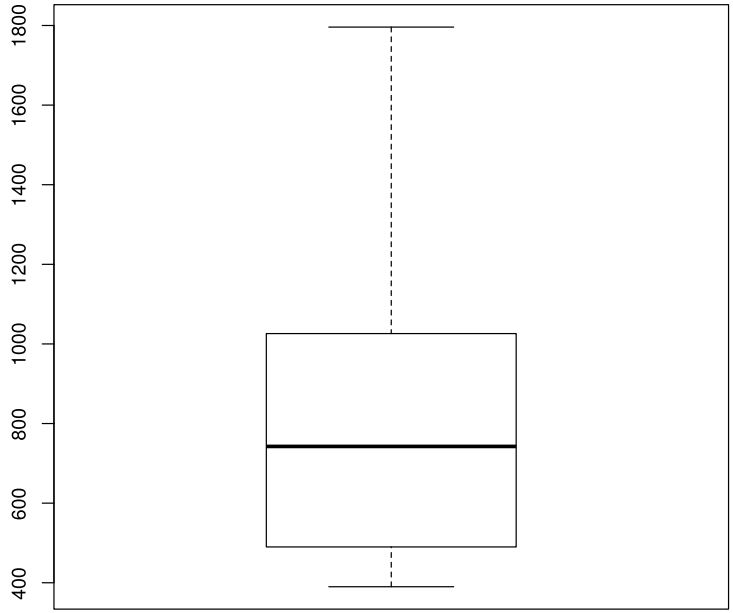
(Sajous et al., 2018)

Mise à jour des dictionnaires : *when (& how much)?*

Fréquence des mises à jour

- *DAF8* : 1932-1935 → *DAF9* : en cours de rédaction
- mises à jour annuelles, *e.g. Petit Larousse* et *Petit Robert*
- *OED* : mises à jour trimestrielles + mises à jour exceptionnelles

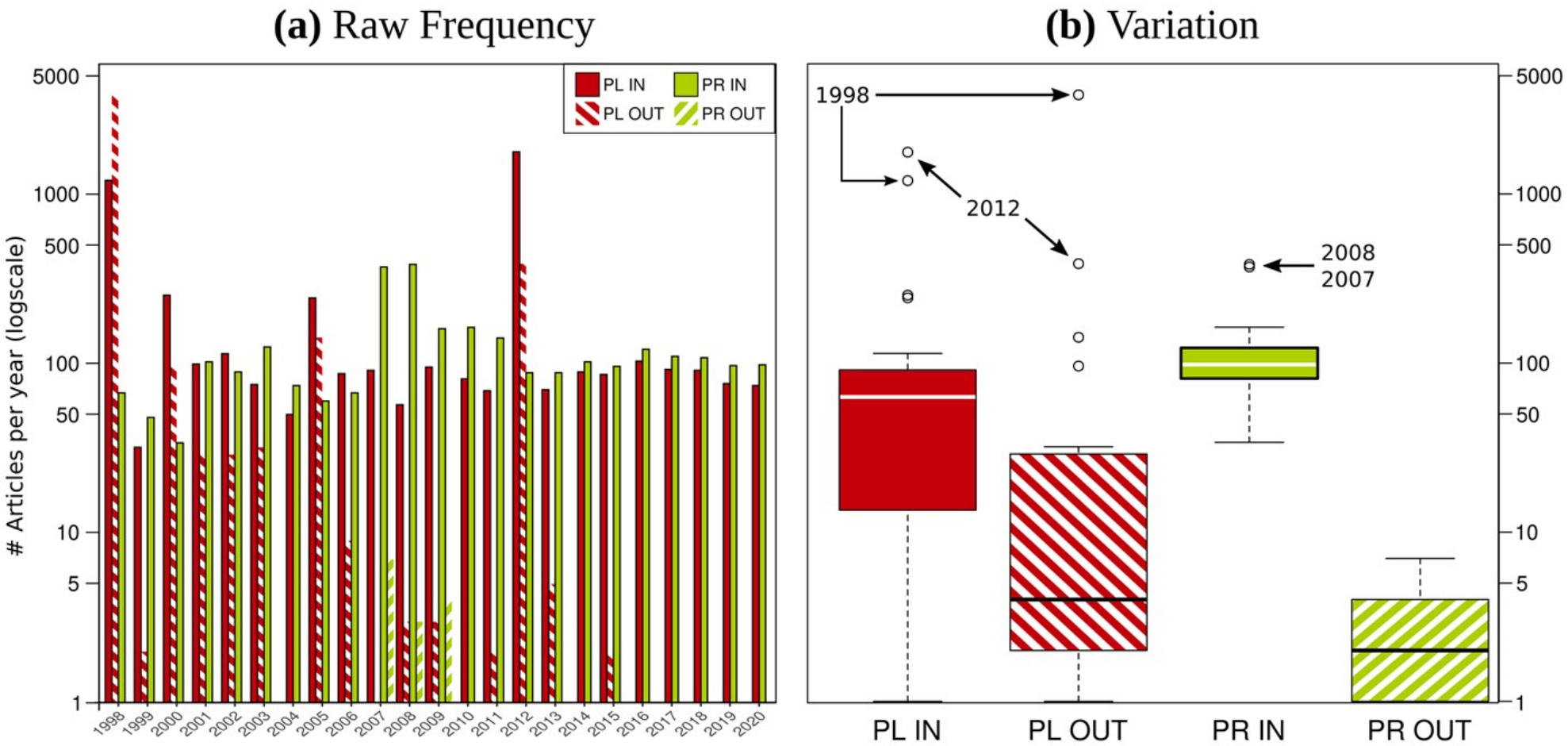
  `https://public.oed.com/updates/`
- mises à jour continue (dictionnaires en ligne), *e.g. Usito*



***OED* : nouveaux articles**	
Min.	390.0
1st Qu.	492.0
Median	742.5
Mean	791.0
3rd Qu.	1014.5
Max.	1796.0

(Sajous et al., 2018)

Mise à jour des dictionnaires : *when (& how much)?*



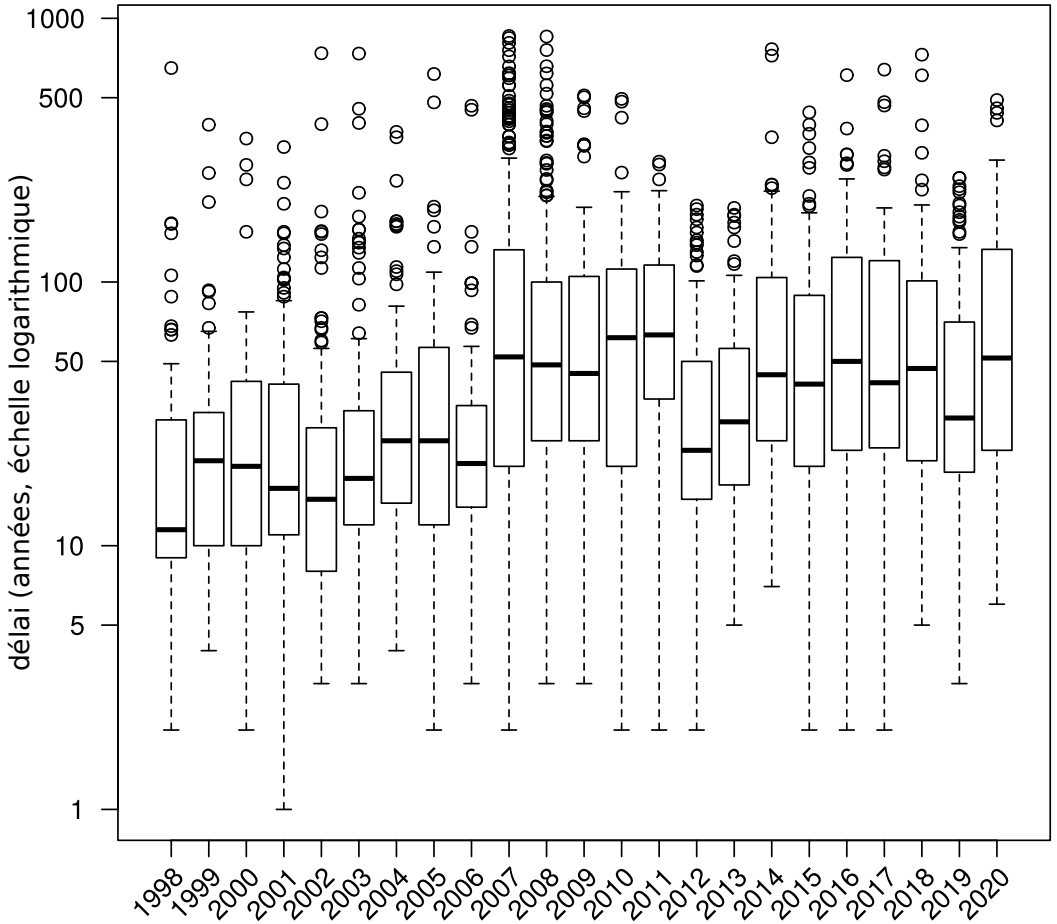
PL : 69 nouveaux articles / millésime, *PR* : 100 nouveaux articles

(Sajous and Martinez, 2021)

Mise à jour des dictionnaires : *when (& how much)?*

Âge des nouvelles entrées

PR : délai première attestation → entrée dans la nomenclature



délai médian

- avant 2007 : 11,5 à 25 ans
- depuis 2007 : 23 à 63 ans

(Sajous and Martinez, 2021)

Mise à jour des dictionnaires : *what & when*

Bilan provisoire

Qu'est-ce qui entre dans le dictionnaire ?

- des néologismes récents, voire très récents (*e.g. Covid*)
- des néologismes installés depuis quelques temps dans l'usage
- des mots de la langue générale existant depuis longtemps (rattrapages)
- des termes techniques parce qu'ils passent dans la langue générale
- quotas de termes techniques de domaines spécialisés au côté de variations diatopiques (selon ligne éditoriale)

N.B. 1

1. nouvelle entrée $\not\Rightarrow$ néologisme
2. néologismes bienvenus

N.B. 2

jusqu'ici, on n'a parlé que d'inclusion de nouvelles entrées

Mise à jour des dictionnaires : *how?*

Lexicographie anglaise/américaine (et à peu près partout ailleurs)

"the corpus revolution" (Rundell and Stock, 1992), "automating the creation of dictionaries" (Rundell and Kilgarriff, 2011) (et réfs par dizaines voire centaines) + contributions participatives (Rundell, 2017) + crowdsourcing (Čibej et al., 2015; Kosem et al., 2013)

Chez Larousse, années 1980 : « chasse aux néologismes » (Sommant, 2000)

- dépouillement manuel quotidien/hebdomadaire de sources diverses
- « flair, sixième sens du lexicographe » (sic)
- comité linguistique $\approx$ 10 personnes, 2 ou 3 séances de travail / an
- catégories et quotas prédéfinis : centaine de néologismes incluant des termes, régionalismes et mots de la francophonie

Mise à jour des dictionnaires : *how?*

Lexicographie anglaise/américaine (et à peu près partout ailleurs)

"the corpus revolution" (Rundell and Stock, 1992), "automating the creation of dictionaries" (Rundell and Kilgarriff, 2011) (et réfs par dizaines voire centaines) + contributions participatives (Rundell, 2017) + crowdsourcing (Čibej et al., 2015; Kosem et al., 2013)

Chez Larousse, années 1980 : « chasse aux néologismes » (Sommant, 2000)

- dépouillement manuel quotidien/hebdomadaire de sources diverses
- « flair, sixième sens du lexicographe » (sic)
- comité linguistique $\approx$ 10 personnes, 2 ou 3 séances de travail / an
- catégories et quotas prédéfinis : centaine de néologismes incluant des termes, régionalismes et mots de la francophonie

Et aujourd'hui (en France) ?

Mise à jour des dictionnaires : *how?*

Lexicographie anglaise/américaine (et à peu près partout ailleurs)

"the corpus revolution" (Rundell and Stock, 1992), "automating the creation of dictionaries" (Rundell and Kilgarriff, 2011) (et réfs par dizaines voire centaines) + contributions participatives (Rundell, 2017) + crowdsourcing (Čibej et al., 2015; Kosem et al., 2013)

Chez Larousse, années 1980 : « chasse aux néologismes » (Sommant, 2000)

- dépouillement manuel quotidien/hebdomadaire de sources diverses
- « flair, sixième sens du lexicographe » (sic)
- comité linguistique $\approx$ 10 personnes, 2 ou 3 séances de travail / an
- catégories et quotas prédéfinis : centaine de néologismes incluant des termes, régionalismes et mots de la francophonie

Et aujourd'hui (en France) ?

Pareil, sans lexicographes ni linguistes

Mise à jour des dictionnaires : comment on sait tout ça ?

Lexicographie anglaise/américaine

- collaboration éditeurs/universitaires
  → manuels par dizaines, articles scientifiques par centaines
- communication vers le public : paratexte + sites et blogs des éditeurs

En France

- pas ou plus de collaboration entre éditeurs et universitaires
- paratexte pauvre, dossiers de presse minces (repris tels quels)

→ sources = enquêtes métalexicographiques et espionnage

Pourquoi les éditeurs français n'utilisent pas de corpus ?

Démarche scientifique *vs.* érudition

- rejet idéologique des corpus par des/un lexicographe(s) influent(s)
- à propos de la démarche philologique du Trésor de la Langue Française, puis de celle du Co-Build :
  « *frénésie scientiste du corpus alimenté par ordinateur que l'on retrouve notamment en Grande-Bretagne* » (Rey, 1995)

Conseils de lecture

- pour une description énamourée de la lexicographie française :
  lire Pruvost (2006)
- pour une description critique de la lexicographie française :
  lire Corbin (1998) et Corbin (2008)

Pourquoi je n'utilise pas de corpus ?

Pourquoi je n'utilise pas de corpus ?

Un corpus ? Où ça ?

Pas de corpus diachronique satisfaisant* du français disponible

*récent, diversifié, mis à jour régulièrement, téléchargeable

Pourquoi je n'utilise pas de corpus ?

Un corpus ? Où ça ?

Pas de corpus diachronique satisfaisant* du français disponible

*récent, diversifié, mis à jour régulièrement, téléchargeable

« Hard parts of lexicography » (Kilgarriff, 1998)

Avec un corpus, c'est [*facile/pas facile/les deux*] de :

- construire la nomenclature d'un dictionnaire :

Pourquoi je n'utilise pas de corpus ?

Un corpus ? Où ça ?

Pas de corpus diachronique satisfaisant* du français disponible

*récent, diversifié, mis à jour régulièrement, téléchargeable

« Hard parts of lexicography » (Kilgarriff, 1998)

Avec un corpus, c'est [*facile/pas facile/les deux*] de :

- construire la nomenclature d'un dictionnaire : facile !
- détecter la néologie formelle :

Pourquoi je n'utilise pas de corpus ?

Un corpus ? Où ça ?

Pas de corpus diachronique satisfaisant* du français disponible

*récent, diversifié, mis à jour régulièrement, téléchargeable

« Hard parts of lexicography » (Kilgarriff, 1998)

Avec un corpus, c'est [*facile/pas facile/les deux*] de :

- construire la nomenclature d'un dictionnaire : facile !
- détecter la néologie formelle : les deux !
  facile : « simple maths » (Kilgarriff, 2009)
  encore que... possible uniquement un certain temps après apparition premières attestations (Falk et al., 2014)

Pourquoi je n'utilise pas de corpus ?

Un corpus ? Où ça ?

Pas de corpus diachronique satisfaisant* du français disponible

*récent, diversifié, mis à jour régulièrement, téléchargeable

« Hard parts of lexicography » (Kilgarriff, 1998)

Avec un corpus, c'est [*facile/pas facile/les deux*] de :

- construire la nomenclature d'un dictionnaire : facile !
- détecter la néologie formelle : les deux !
  facile : « simple maths » (Kilgarriff, 2009)
  encore que... possible uniquement un certain temps après apparition premières attestations (Falk et al., 2014)
- détecter la néologie sémantique :

Pourquoi je n'utilise pas de corpus ?

Un corpus ? Où ça ?

Pas de corpus diachronique satisfaisant* du français disponible

*récent, diversifié, mis à jour régulièrement, téléchargeable

« Hard parts of lexicography » (Kilgarriff, 1998)

Avec un corpus, c'est [*facile/pas facile/les deux*] de :

- construire la nomenclature d'un dictionnaire : facile !
- détecter la néologie formelle : les deux !
  facile : « simple maths » (Kilgarriff, 2009)
  encore que... possible uniquement un certain temps après apparition premières attestations (Falk et al., 2014)
- détecter la néologie sémantique : pas facile !
  clustering non supervisé (Cook et al., 2013)

Pourquoi je n'utilise pas de corpus ?

Un corpus ? Où ça ?

Pas de corpus diachronique satisfaisant* du français disponible

*récent, diversifié, mis à jour régulièrement, téléchargeable

« Hard parts of lexicography » (Kilgarriff, 1998)

Avec un corpus, c'est [*facile/pas facile/les deux*] de :

- construire la nomenclature d'un dictionnaire : facile !
- détecter la néologie formelle : les deux !
  facile : « simple maths » (Kilgarriff, 2009)
  encore que... possible uniquement un certain temps après apparition premières attestations (Falk et al., 2014)
- détecter la néologie sémantique : pas facile !
  clustering non supervisé (Cook et al., 2013)
- oui, mais aujourd'hui avec les embeddings...

Pourquoi je n'utilise pas de corpus ?

Un corpus ? Où ça ?

Pas de corpus diachronique satisfaisant* du français disponible

*récent, diversifié, mis à jour régulièrement, téléchargeable

« Hard parts of lexicography » (Kilgarriff, 1998)

Avec un corpus, c'est [*facile/pas facile/les deux*] de :

- construire la nomenclature d'un dictionnaire : facile !
- détecter la néologie formelle : les deux !
  facile : « simple maths » (Kilgarriff, 2009)
  encore que... possible uniquement un certain temps après apparition premières attestations (Falk et al., 2014)
- détecter la néologie sémantique : pas facile !
  clustering non supervisé (Cook et al., 2013)
- oui, mais aujourd'hui avec les embeddings... pas facile quand même !
  - détection des changements sémantiques : périodes longues (décennies ou siècles), très gros corpus (Kutuzov et al., 2018)
  - re-un-corpus-où-ça ?

Une proposition non exclusive

Méthode proposée : palliatif ou démarche complémentaire

- palliatif quand corpus indisponible, complément le reste du temps
- option idéale = analyse de corpus + "good old-fashioned lexicography" (Rundell, 2002) + log de Wiktionary + ...

Conditions de productions d'un dictionnaire

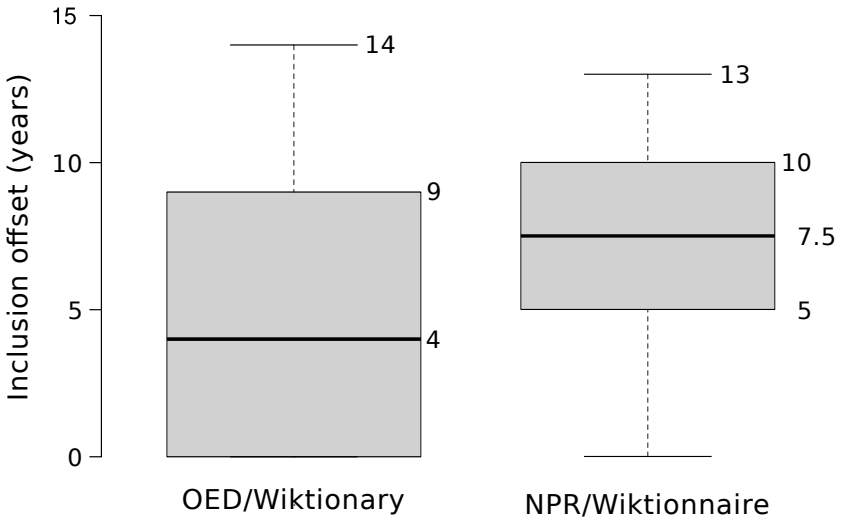
"Dictionaries are not written in a vacuum, but by people working under the pressure of time. It sometimes seems to me that as technology has improved the speed and power with which we can examine the language, the pressures to produce quickly and with fewer staff have kept pace, so that on balance nothing is accomplished any faster or better. The expectations of management seem to rise at the same rate as the speed and power of the computer increase [...] Corpora can be used well or they can be used badly. Time pressures too often push the lexicographer to cut corners to avoid time-consuming analyses. It really doesn't do much good having a good corpus with marvelous analytical tools if they aren't used." (Landau, 2001, p. 323)

Wiktionary, source pour la détection de créations lexicales

Nouvelles entrées 2017 dictionnaires commerciaux / présence dans Wiktionary (Sajous et al., 2018)

NPR **entries**		***Wiktionnaire***	
		Present	Missing
monolex	112	104 (93 %)	8 (7 %)
polylex	36	12 (33 %)	24 (67 %)
TOTAL	163	128 (79 %)	35 (21 %)

OED **entries**		***Wiktionary***	
		Present	Missing
monolex	264	198 (75 %)	66 (25 %)
polylex	55	34 (62 %)	21 (38 %)
TOTAL	319	232 (73 %)	87 (27 %)



Historique des révisions de Wiktionary...

∃ méthodes plus simples : liste de mots prêtes à l'emploi

- Catégorie *Coronavirus* : 54 articles (01/2020) → 124 articles (06/2020)
  (`https://en.wiktionary.org/wiki/Category:en:Coronavirus`)
- Catégorie *Hot words newer than a year*
  (`https://en.wiktionary.org/wiki/Category:Hot_words_newer_than_a_year`)
  - 01/2021 : 94 mots, dont 26 ne sont pas anglais.
  - 79% des mots anglais (54/68) sont liés au Covid

→ dépendantes de l'édition de langue (*e.g.* pas de catégories équivalentes en français), et du sujet (+ nombreux mots pertinents non catégorisés)

Recherche patrons dans vedettes & définitions (*e.g. corona*, *covid*)

- vedettes : ***covid****iot*, ***covid*** *party*, ***corona****sceptic*, ***corona****viruslike*...
- définitions : *long hauleur = a* ***COVID****-19 patient who is suffering from [...]*

social distancing = The practice of maintaining physical distance between people to reduce the spread of communicable diseases

Historique des révisions de Wiktionary...

Dump historique de Wiktionary (https://dumps.wikimedia.org/)

- contient toutes les versions de tous les articles
- une révision → timestamp + contributeur

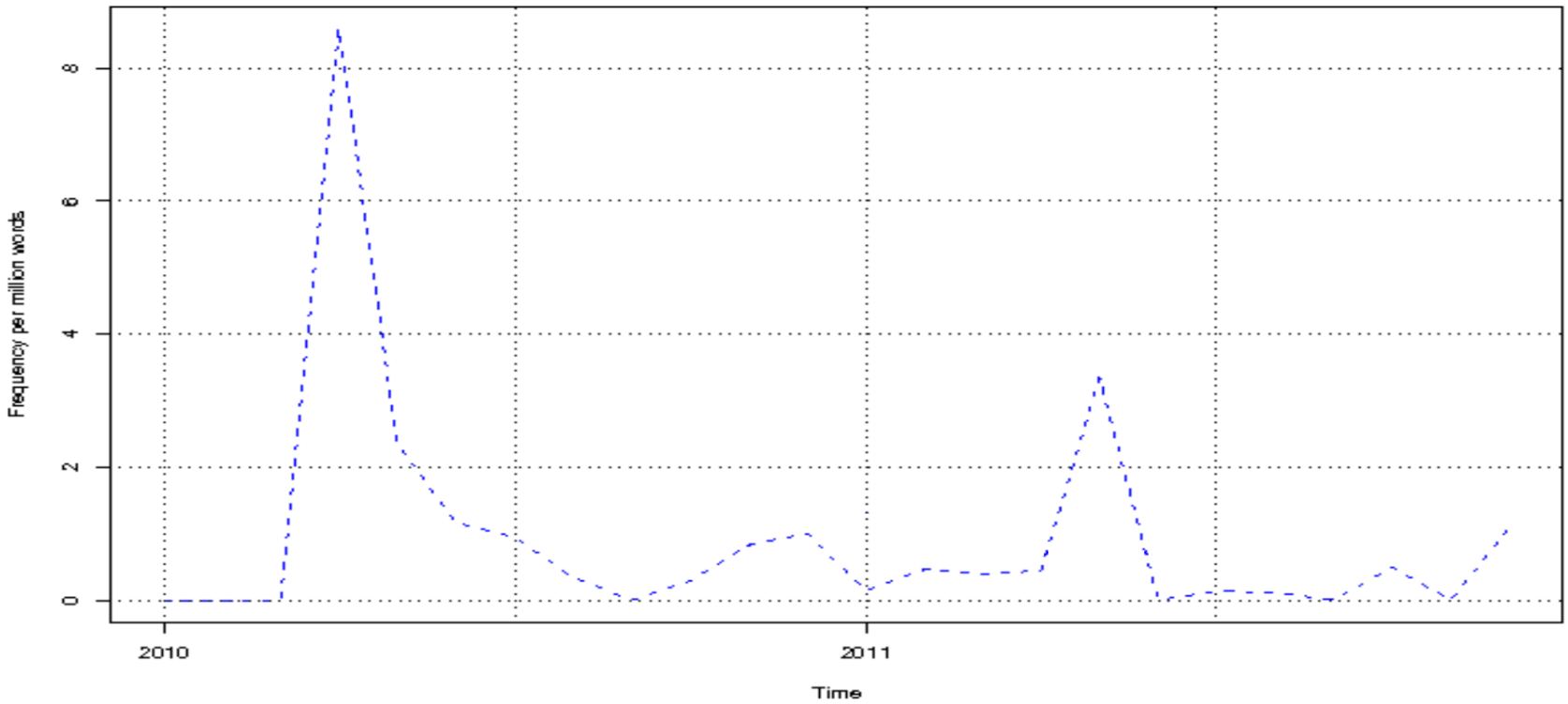
Travaux précédents sur les logs de Wikipédia & Wiktionary

- Lih (2004) : qualité des articles de la Wikipedia anglaise
  - « rigueur » : nb révisions/article
  - « diversité » : nb contributeurs/article
- Wolfer and Müller-Spitzer (2016) :
  dynamique des éditions anglaise et allemande de Wiktionary
  - nb révisions/article corrélé avec fréquence des mots en corpus

Historique des révisions de Wiktionary...

Retour au corpus

- Renouf (2013) : "lifecycle of words"
  - fluctuation des fréquences en corpus → processus néologiques



Eyjafjallajökull

Historique des révisions de Wiktionary...

Hypothèses

1. variation importante du nb de révisions → indice de néologie
2. diversité de contributeurs → sujet d'actualité

Néologie, Wiktionary et mise à jour des dictionnaires

- nombreuses révisions dans nouvel article de Wiktionary
  → possible création lexicale
  → candidat à l'ajout au dictionnaire
- nombreuses révisions dans article existant dans Wiktionary
  → possible néologie sémantique
  → ajout de sens dans le dictionnaire ou article à réviser

Corpus

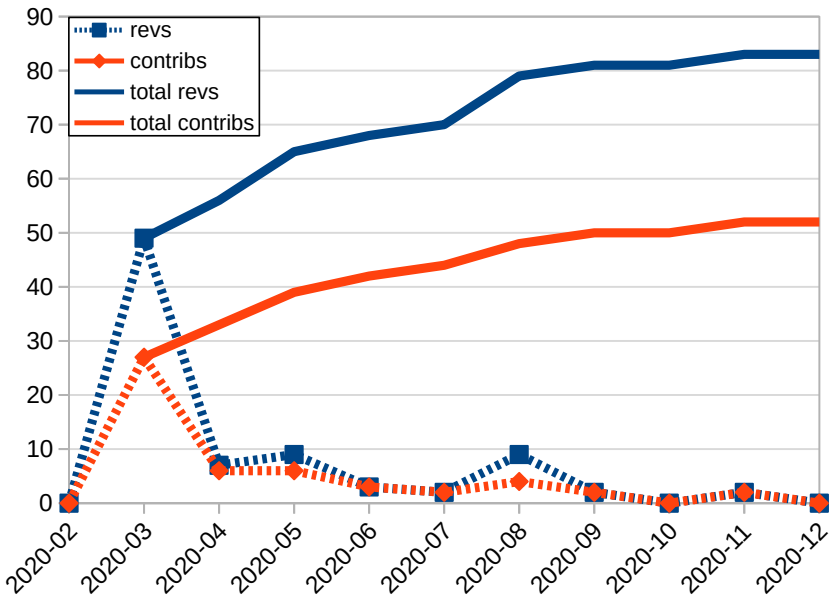
Dump historique de Wiktionary : prétraitements

- Version EN+FR du 1/01/2021
- Extraction pour chaque entrée et chaque révision : date + contributeur
- Filtrage : pages de discussion, pages utilisateurs, mots non lexicaux, flexions, mots d'une autre langue
- Révisions modifiant uniquement une section d'une autre langue ignorées, *e.g.* rév du 5/11/2020, article *coronavirus* de Wiktionary (EN) : ajout du dérivé *coronaviraal* à la section néerlandaise de l'article

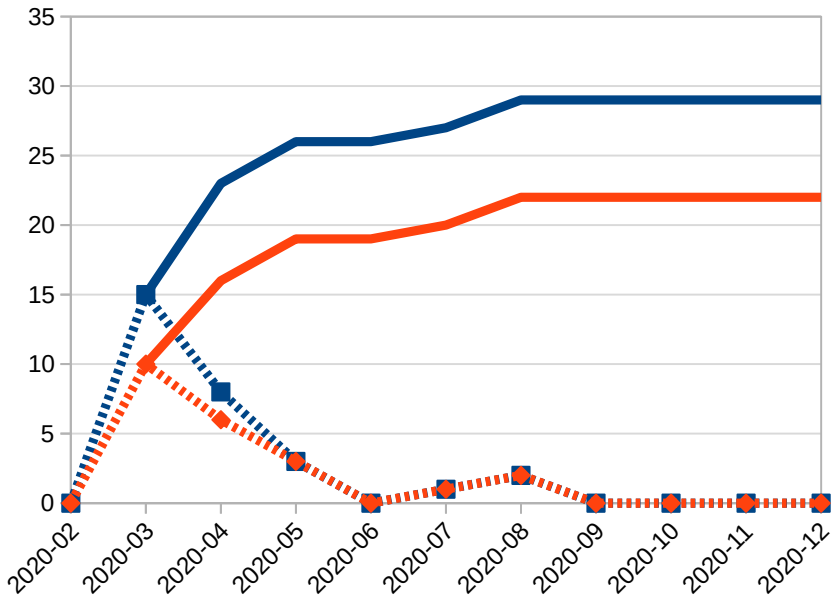
Corpus résultant

- FR : $\approx$ 6 millions de révisions pour 331 000 articles
  11 600 nouveaux lemmes ajoutés en 2020
- EN : $\approx$ 6,4 millions de révisions pour 447 000 articles
  31 000 nouveaux lemmes ajoutés en 2020
- 11 600 & 31 000 créations lexicales + 331 000 & 447 000 néologismes sémantiques potentiels $\Rightarrow$ nécessité de classement par score de pertinence

Nouvelles entrées : nombre de contributions/contributeurs

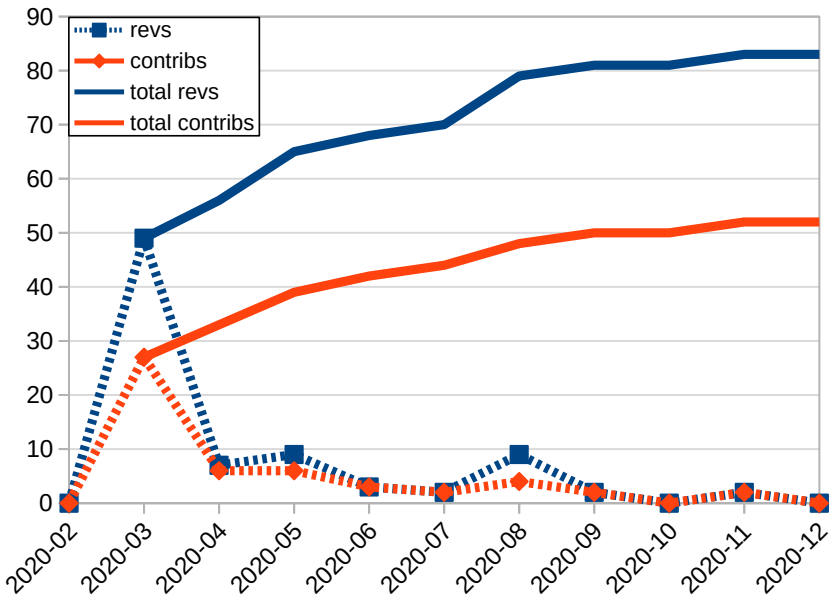


social distancing

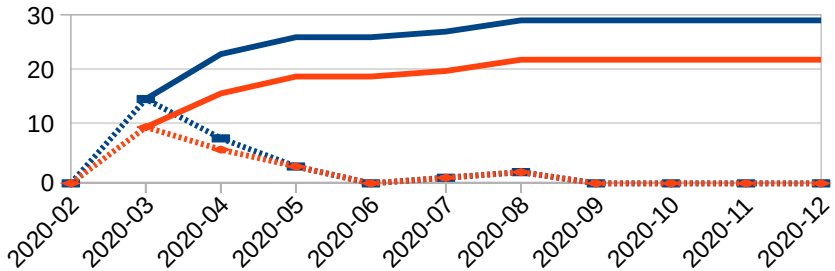


flatten the curve

Nouvelles entrées : nombre de contributions/contributeurs

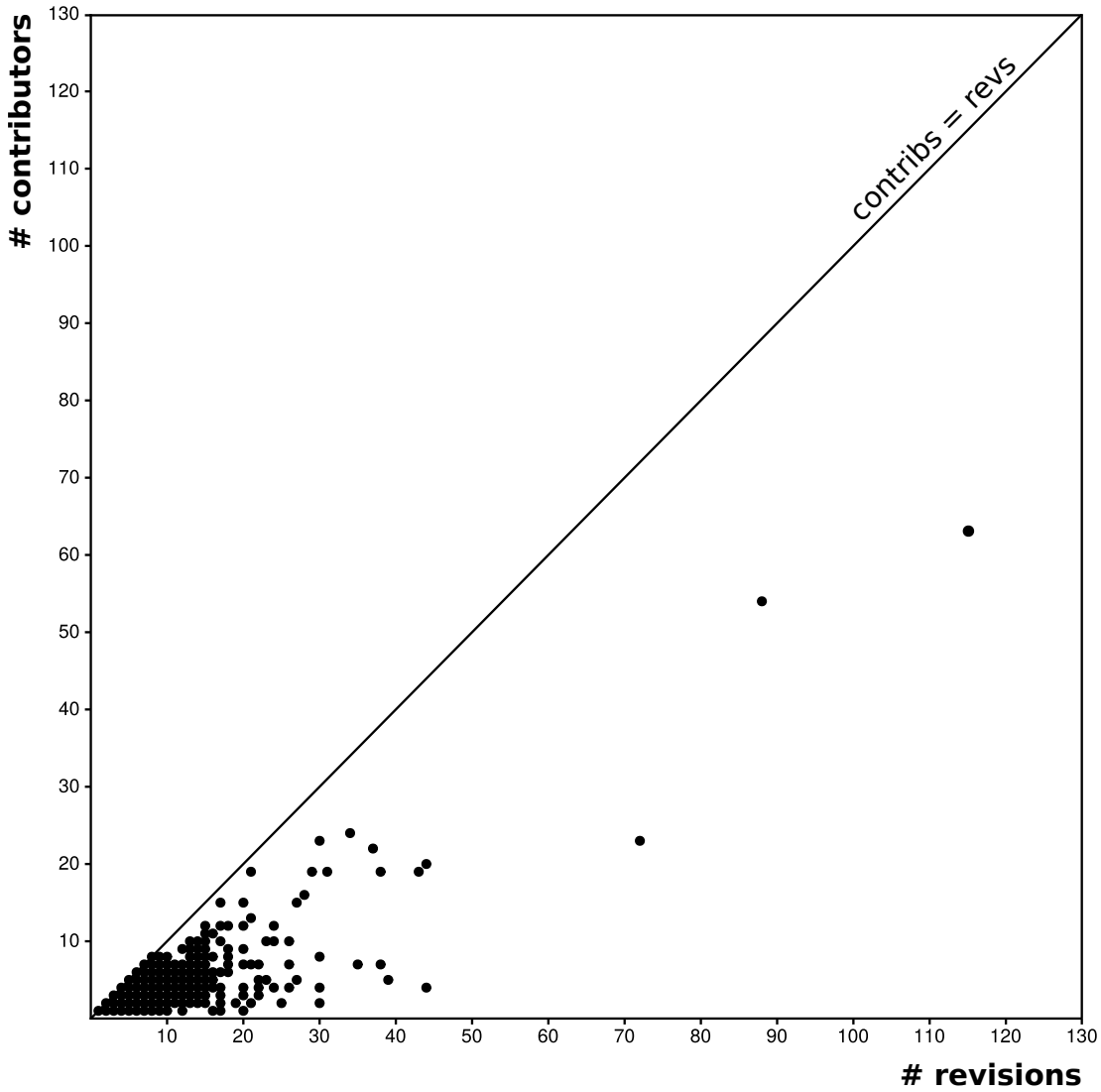


social distancing



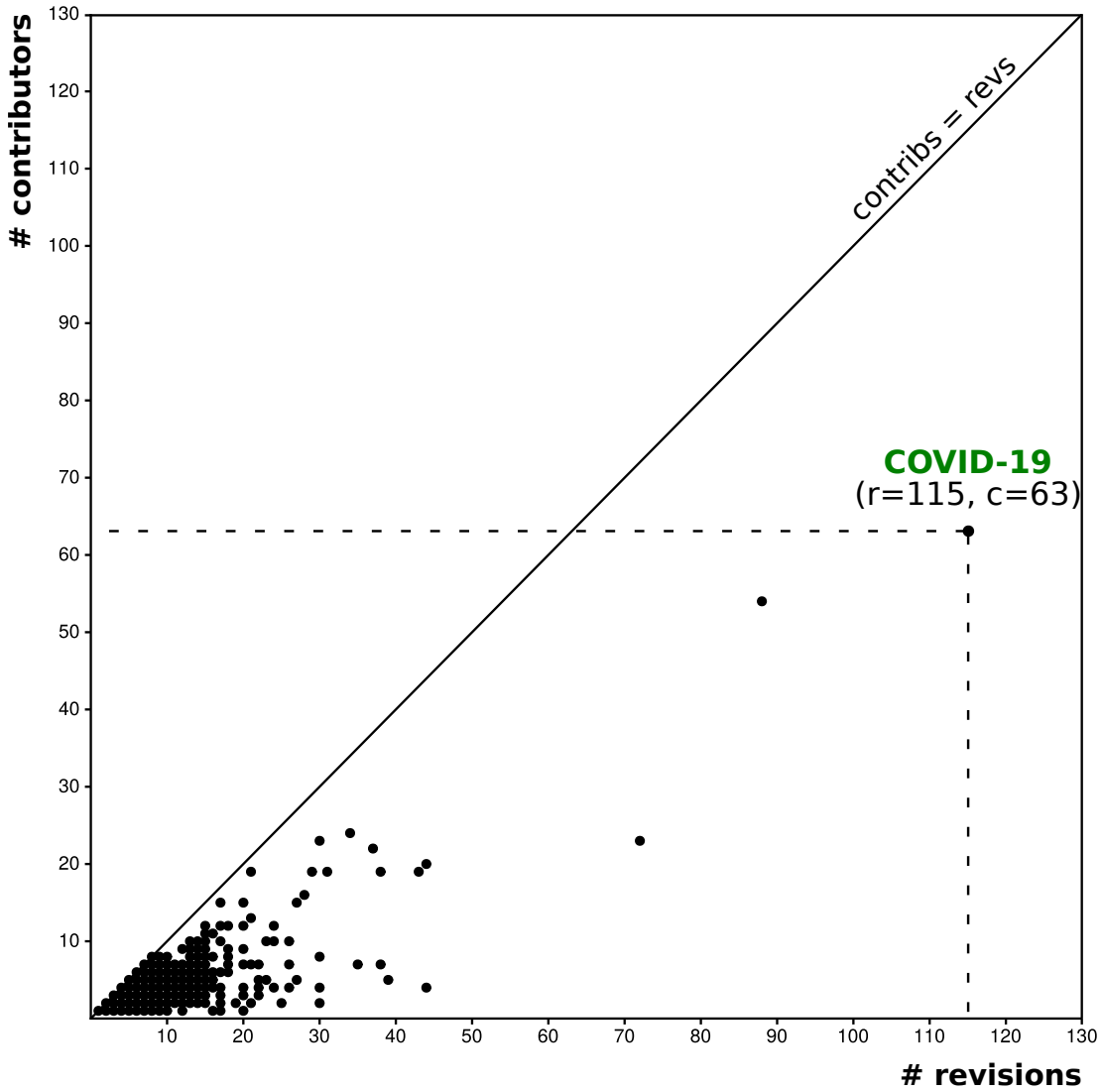
flatten the curve

Diagramme de dispersion : 31 107 **nouveaux** lemmes (EN)



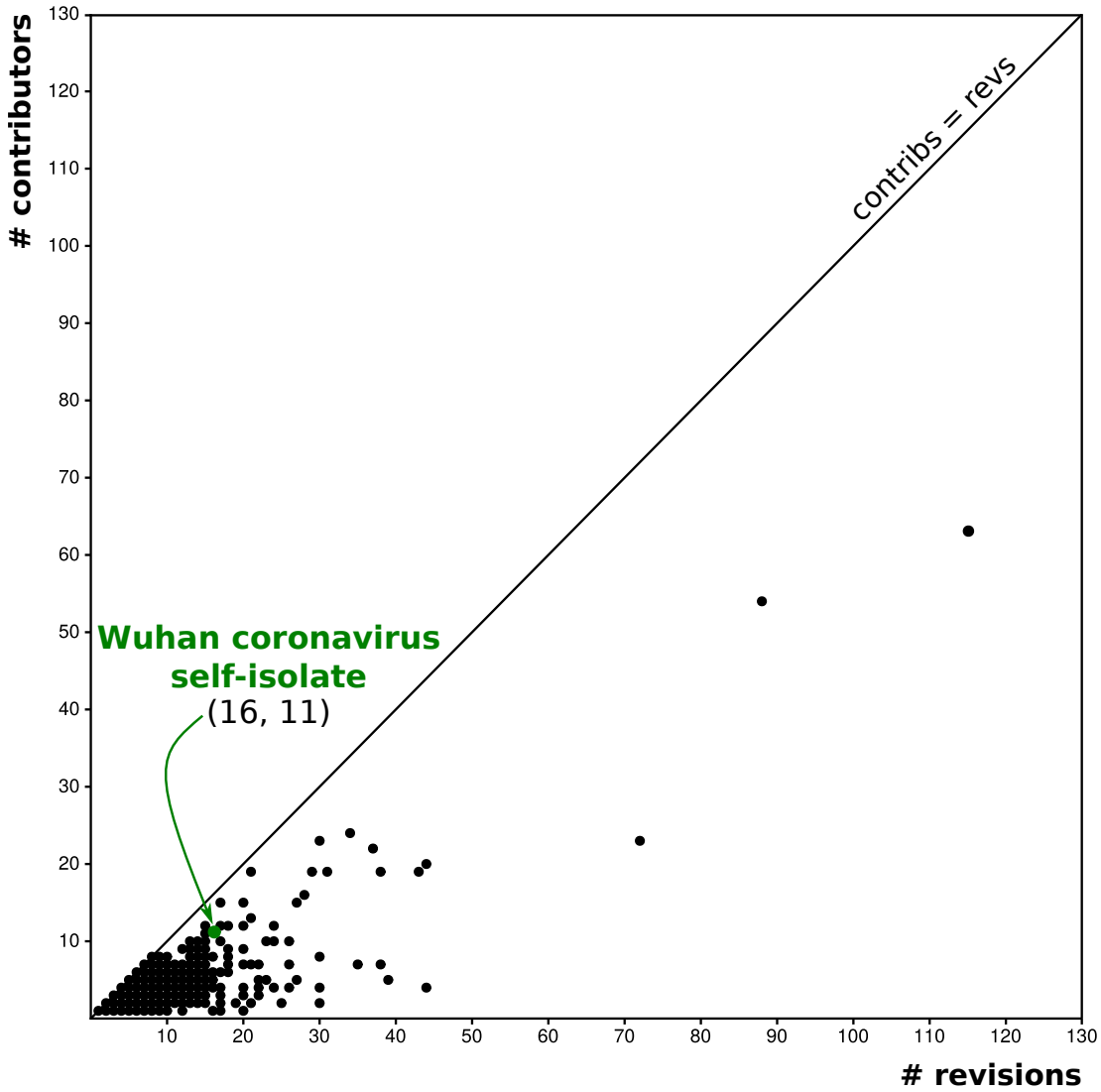
- coordonnées :
  x = # révisions,
  y = # contributeurs

Diagramme de dispersion : 31 107 **nouveaux** lemmes (EN)



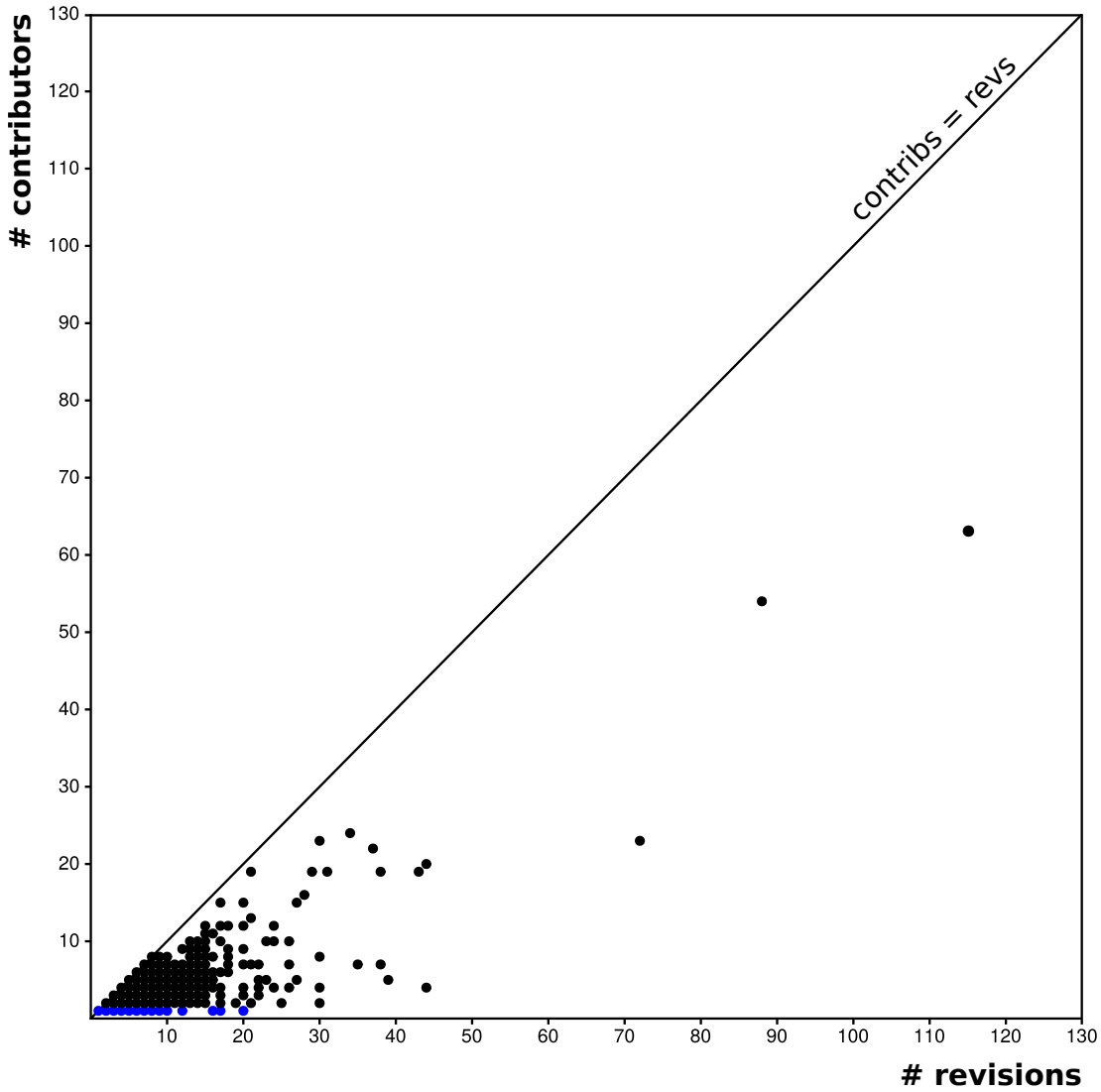
- coordonnées :
  x = # révisions,
  y = # contributeurs
- 1 entrée → 1 point de coordonnées

Diagramme de dispersion : 31 107 **nouveaux** lemmes (EN)



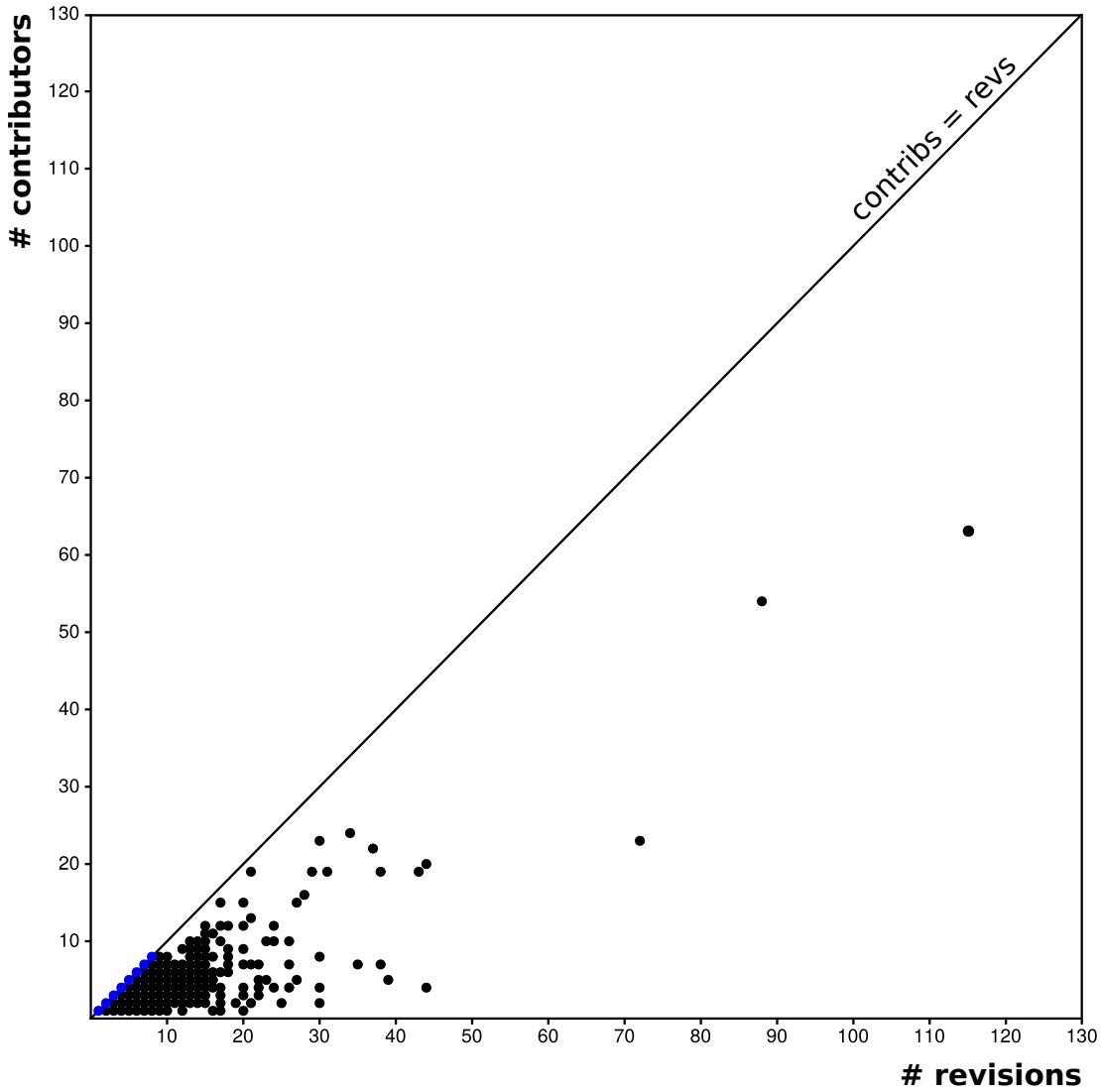
- coordonnées :
  x = # révisions,
  y = # contributeurs
- 1 entrée → 1 point de coordonnées
- 1 point de coordonnées → possiblement plusieurs entrées

Diagramme de dispersion : 31 107 **nouveaux** lemmes (EN)



- coordonnées :
  x = # révisions,
  y = # contributeurs
- 1 entrée → 1 point de coordonnées
- 1 point de coordonnées → possiblement plusieurs entrées
- le long des abscisses → articles édités par un seul contributeur

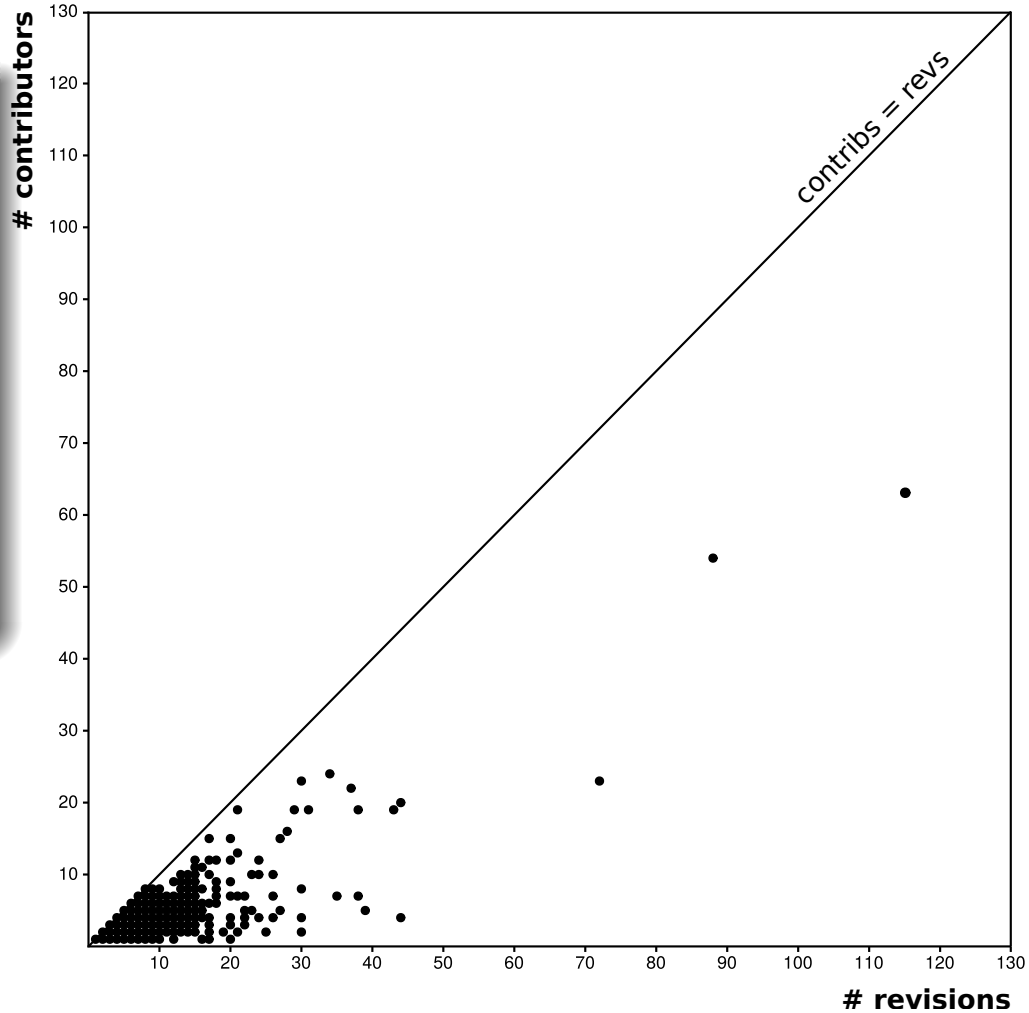
Diagramme de dispersion : 31 107 **nouveaux** lemmes (EN)



- coordonnées :
  x = # révisions,
  y = # contributeurs
- 1 entrée → 1 point de coordonnées
- 1 point de coordonnées → possiblement plusieurs entrées
- le long des abscisses → articles édités par un seul contributeur
- le long de la diagonale (#contribs = #revs) → chaque contributeur a édité l'article une seule fois

Classement des **nouvelles** entrées

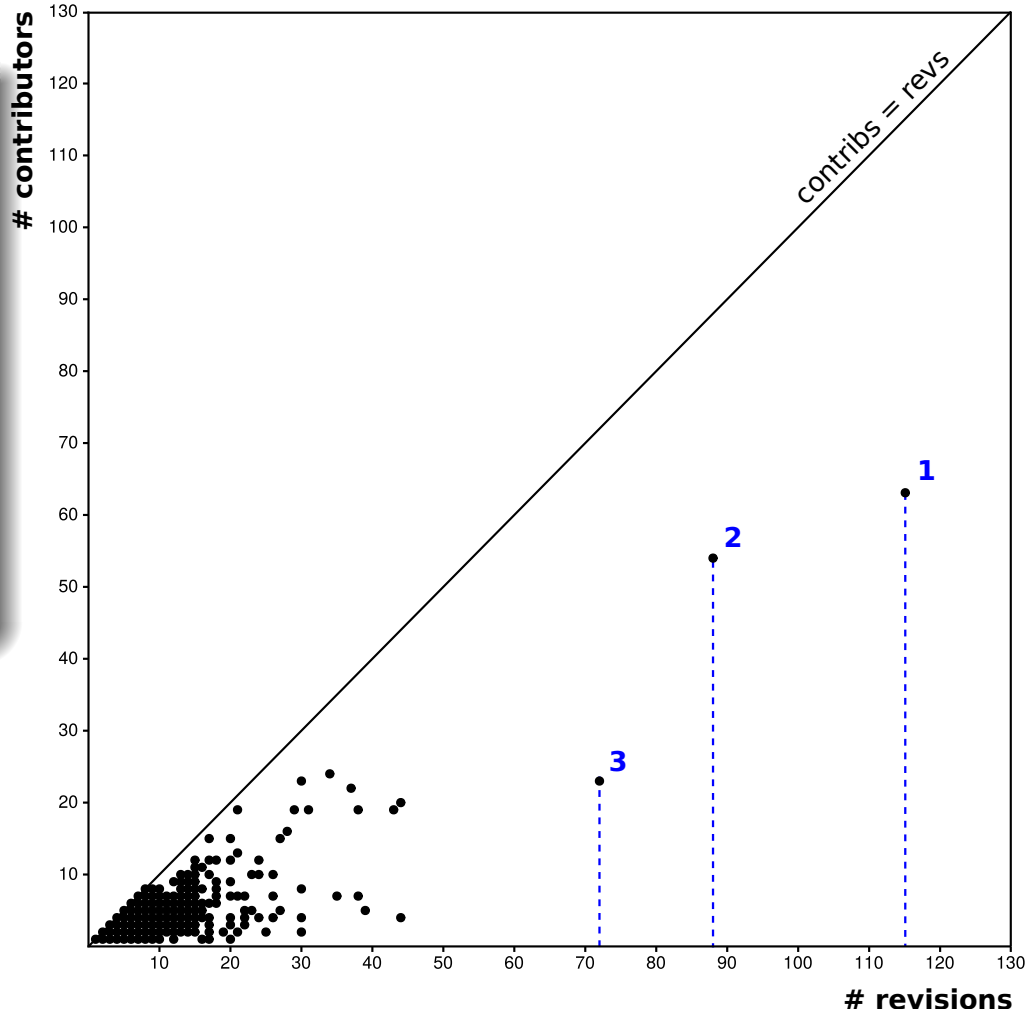
Scores



Classement des **nouvelles** entrées

Scores

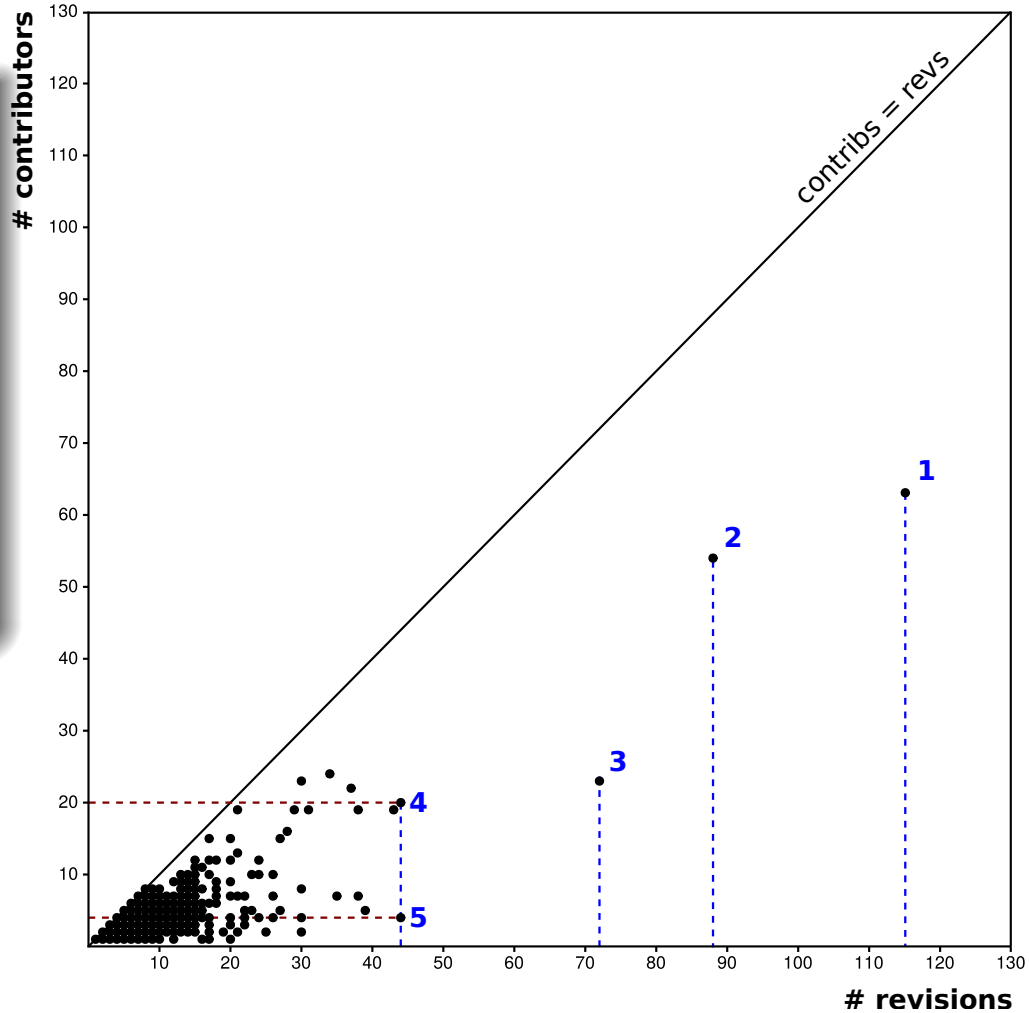
- # révisions



Classement des **nouvelles** entrées

Scores

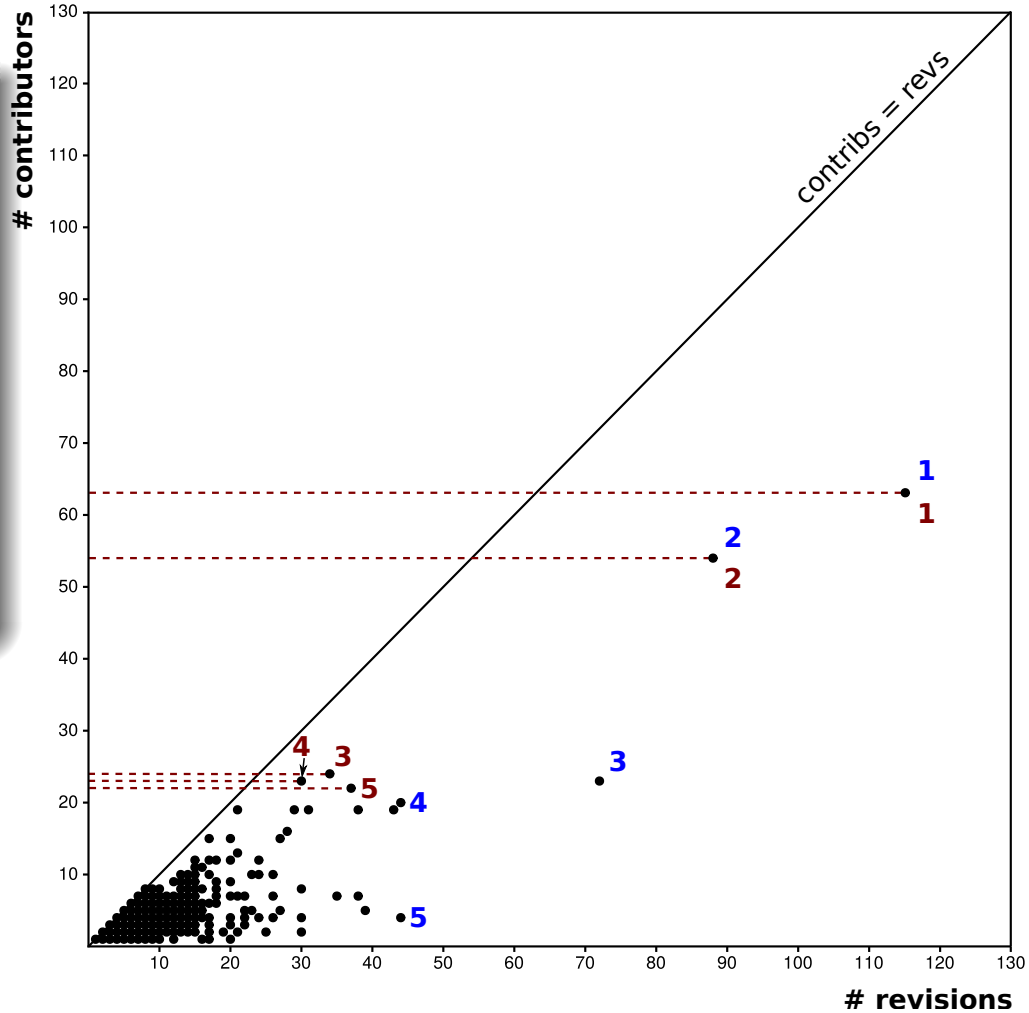
- # révisions



Classement des **nouvelles** entrées

Scores

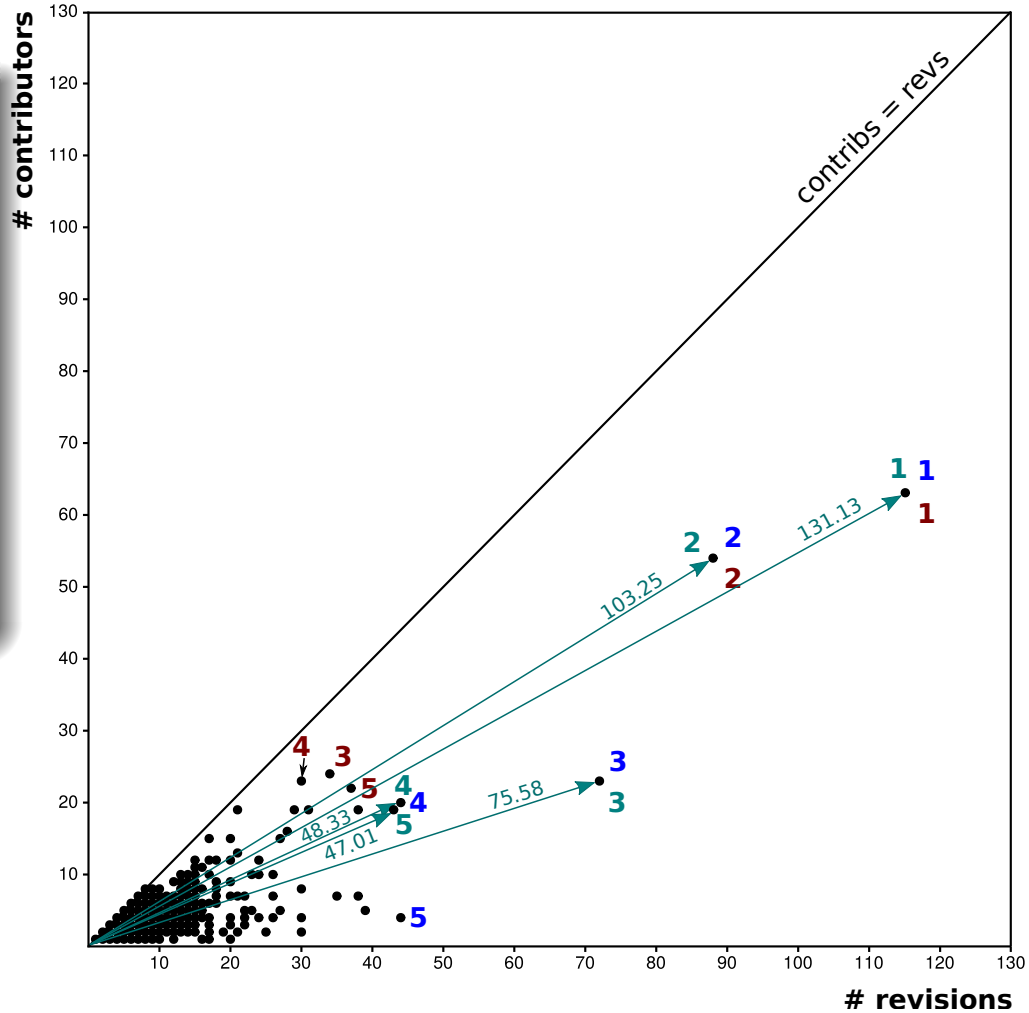
- # révisions
- # contributeurs distincts



Classement des **nouvelles** entrées

Scores

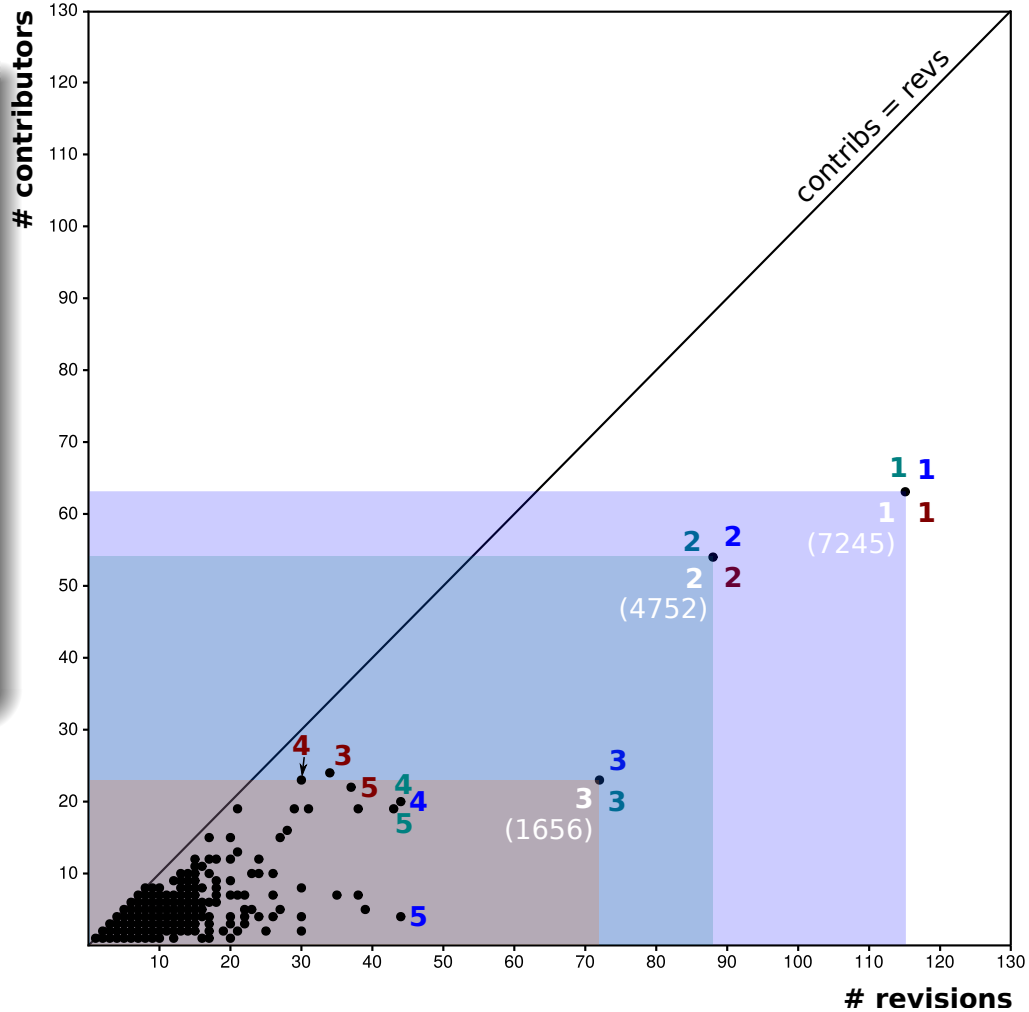
- # révisions
- # contributeurs distincts
- distance : $\sqrt{revs^2 + contribs^2}$



Classement des **nouvelles** entrées

Scores

- # révisions
- # contributeurs distincts
- distance : $\sqrt{revs^2 + contribs^2}$
- produit : $revs \times contribs$

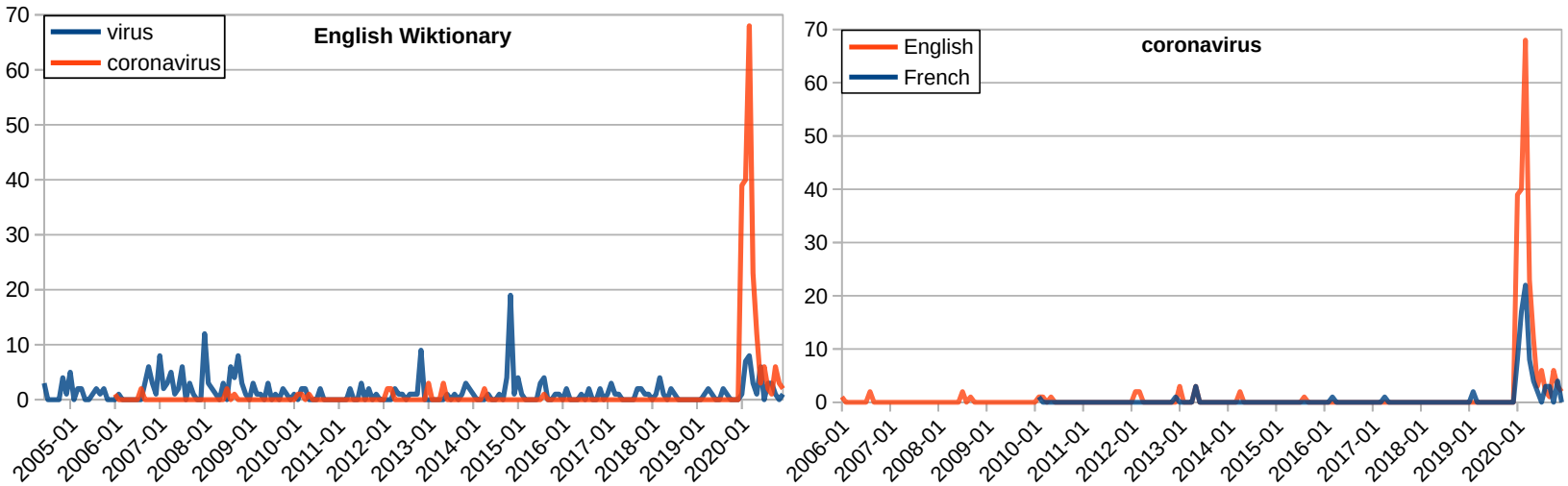


Classement des entrées **existantes**

= détecter des écarts locaux par rapport aux taux de révisions *habituels*

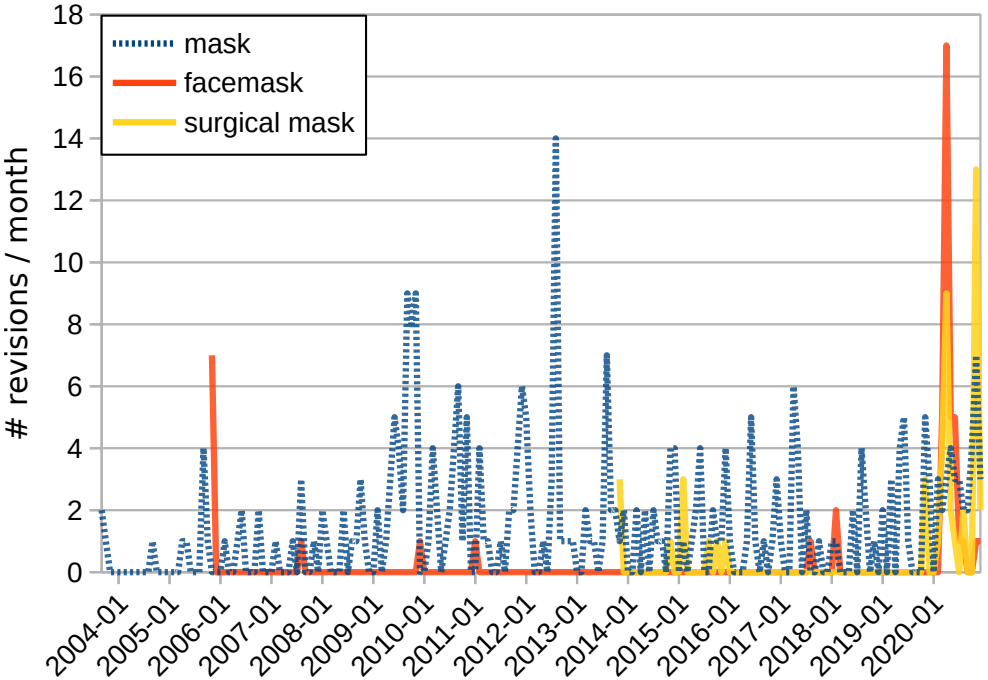
Nb de révisions d'une entrée dépend de ses caractéristiques linguistiques (interdépendantes) et de l'édition de langue

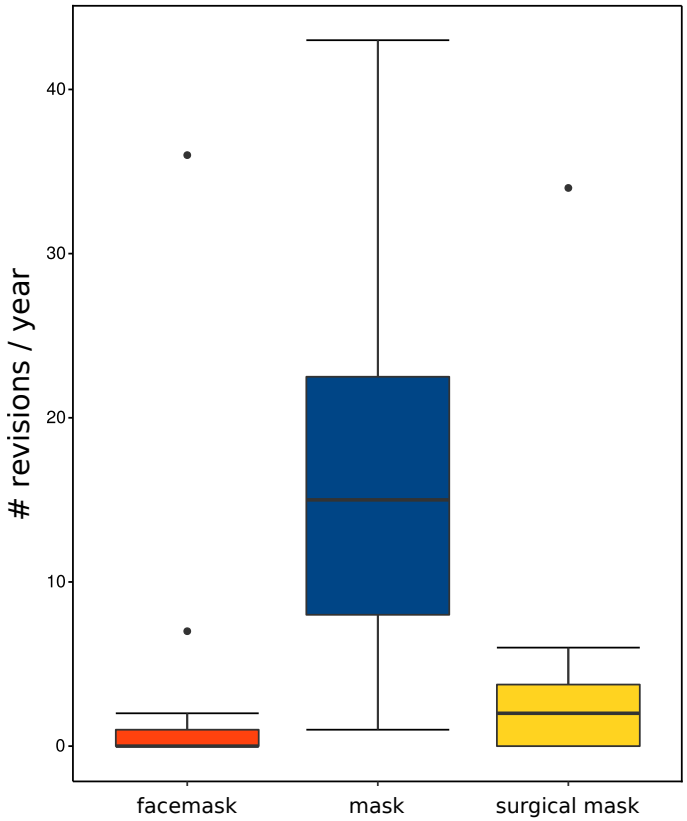
1. fréquence (mot fréquent *vs.* rare)
2. polysémie (mot monosémique *vs.* polysémique)
3. degré de technicité (langue générale *vs.* spécialisée)



Classement des entrées **existantes**

= détecter des écarts locaux par rapport aux taux de révisions *habituels*





Classement des entrées **existantes**

= détecter des écarts locaux par rapport aux taux de révisions *habituels*

Scores

Étant donné une entrée h et une période p (année, trimestre, ...):

1. $avgRevsRatio_p(h) = \frac{1+revs_p(h)}{1+avg(revs(h))}$
2. $medianRevsRatio_p(h) = \frac{1+revs_p(h)}{1+median(revs(h))}$
3. $avgContribsRatio_p(h) = \frac{1+contribs_p(h)}{1+avg(contribs(h))}$
4. $medianContribsRatio_p(h) = \frac{1+contribs_p(h)}{1+median(contribs(h))}$

Annotation des néologismes candidats

Objectifs

- évaluation globale de la méthode :
  est-ce que ça marche (un peu/pas du tout) ?
- comparaison des résultats obtenus avec les différents scores, sur différentes périodes

Annotation

- question : *mot en lien avec le Covid ?* (oui/non)
- 200 premières **nouvelles** entrées
  + 100 premières entrées **existantes** pour :
    - 2 langues
    - 4 scores
    - 5 périodes (année 2020 + 4 trimestres)
    - 4 sources de données (avec/sans robots, avec/sans anonymes)

$\rightarrow$ + de 3 000 entrées anglaises et 3 000 entrées françaises annotées

Annotation des néologismes candidats

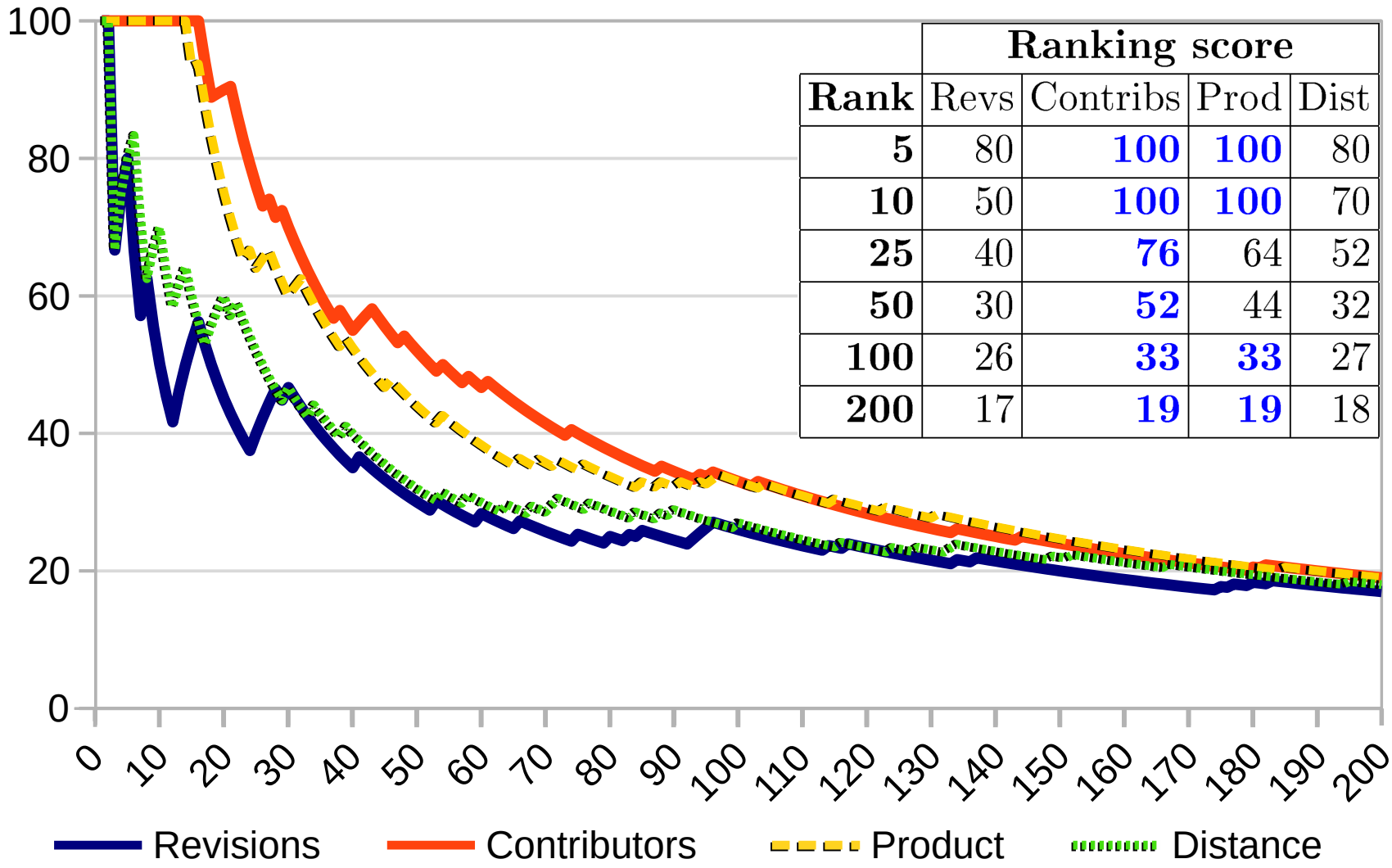
« mot en lien (direct ou raisonnablement indirect) avec le Covid »

- virus/maladie, soin, contrôle de la propagation de la pandémie, stats, conséquence de la pandémie sur les activités professionnelles et sociales...
- Wiktionary : nombreux régionalismes, occasionalismes, mots datés, etc.
- lecture de la définition souvent nécessaire
- recherche de connaissances encyclopédiques parfois nécessaires (noms de médicaments, molécules, etc.) *e.g. pamaquine* (ajout 2009) *quinium* (ajout 2020) : lien avec la *chloroquine* (non mentionnée dans la déf.) *vs. bambuterol* : médicament pour le traitement de l'asthme $\rightarrow$?
  en l'absence de lien clair, annotation négative
- *extractor fan* $\rightarrow$ 0 *vs. ventilator* $\rightarrow$ 1 (= *medical ventilator* $\approx$ *respirator*)
- *doomscrolling* (EN) = "the practice of continually reading Internet news about catastrophic events" $\rightarrow$? [$\in$ catégorie *Coronavirus*]
- *doomscrolling* (FR) = "consommation compulsive d'informations sur Internet, en particulier relatant des catastrophes ou des faits divers d'actualité" $\rightarrow$? [Déjà annoté pour l'anglais !]

20 **nouvelles** entrées les mieux classées

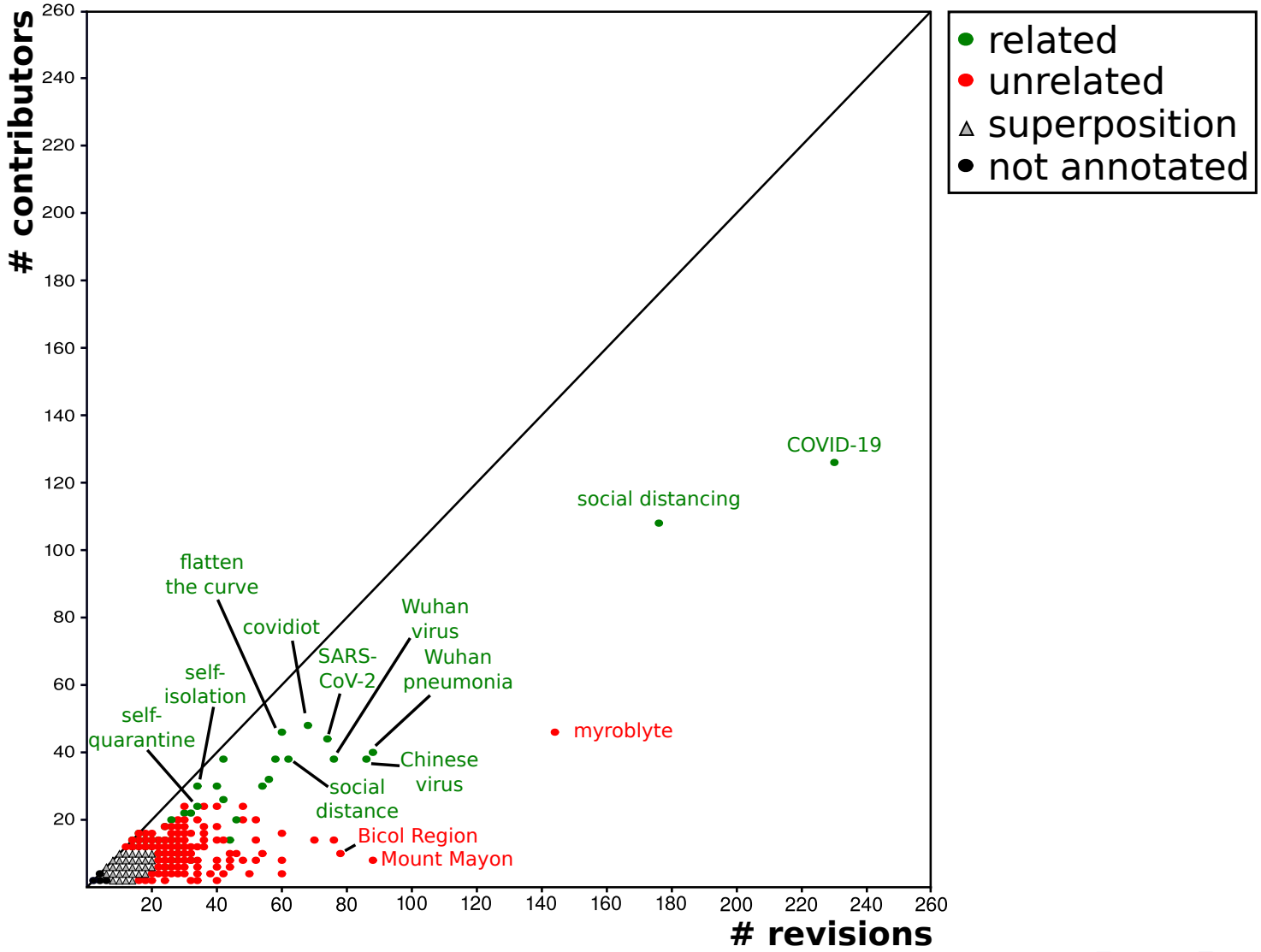
	Ranking score			
Rank	**# Revisions**	**# Contributors**	**Product**	**Distance**
1	COVID-19	COVID-19	COVID-19	COVID-19
2	social distancing	social distancing	social distancing	social distancing
3	Mount Mayon	covidiot	Wuhan pneumonia	Mount Mayon
4	Wuhan pneumonia	flatten the curve	Wuhan virus	Wuhan pneumonia
5	Wuhan virus	Wuhan virus	covidiot	Wuhan virus
6	Bicol Region	Wuhan pneumonia	flatten the curve	Chinese virus
7	Peja	covid	Chinese virus	Bicol Region
8	Chinese virus	social distance	covid	Peja
9	Berat	SARS-CoV-2	infectious disease specialist	covidiot
10	Medusavirus	Chinese virus	social distance	flatten the curve
11	Vlorë	infectious disease specialist	SARS-CoV-2	Berat
12	Kizilsu	contact tracing	contact tracing	Medusavirus
13	covidiot	self-isolation	Kung Flu	covid
14	infectious disease specialist	Kung Flu	Wuhan flu	infectious disease specialist
15	flatten the curve	Wuhan flu	Medusavirus	Vlorë
16	covid	COVID	self-isolation	Kizilsu
17	world map	Spleef	world map	world map
18	Accompong	Trumpster fire	Bicol Region	social distance
19	Arlberg	self-quarantine	Mount Mayon	SARS-CoV-2
20	sinoatrial node	Wuhan coronavirus	wokefish	contact tracing

200 **nouvelles** entrées les mieux classées

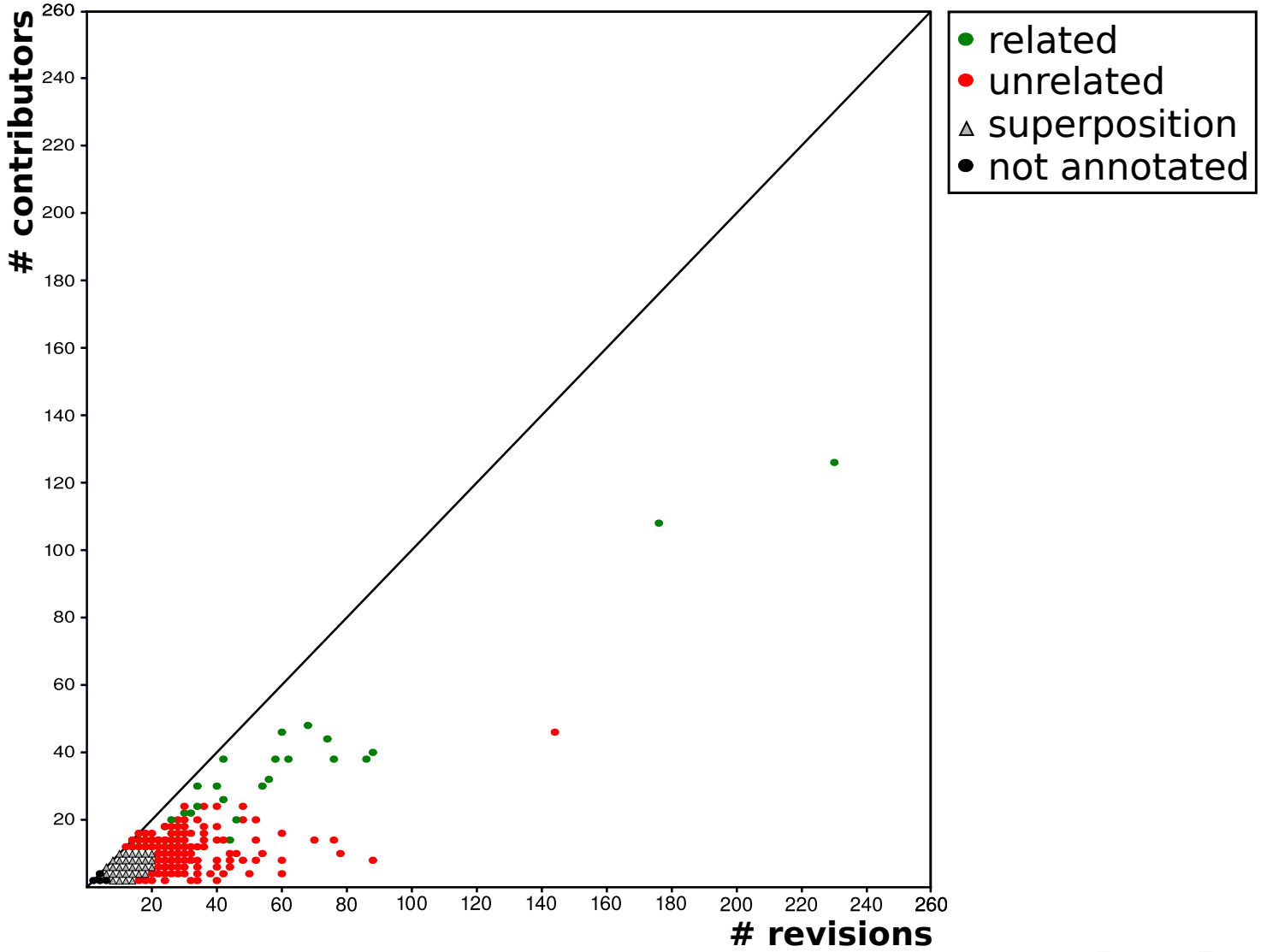


	Ranking score			
Rank	Revs	Contribs	Prod	Dist
5	80	**100**	**100**	80
10	50	**100**	**100**	70
25	40	**76**	64	52
50	30	**52**	44	32
100	26	**33**	**33**	27
200	17	**19**	**19**	18

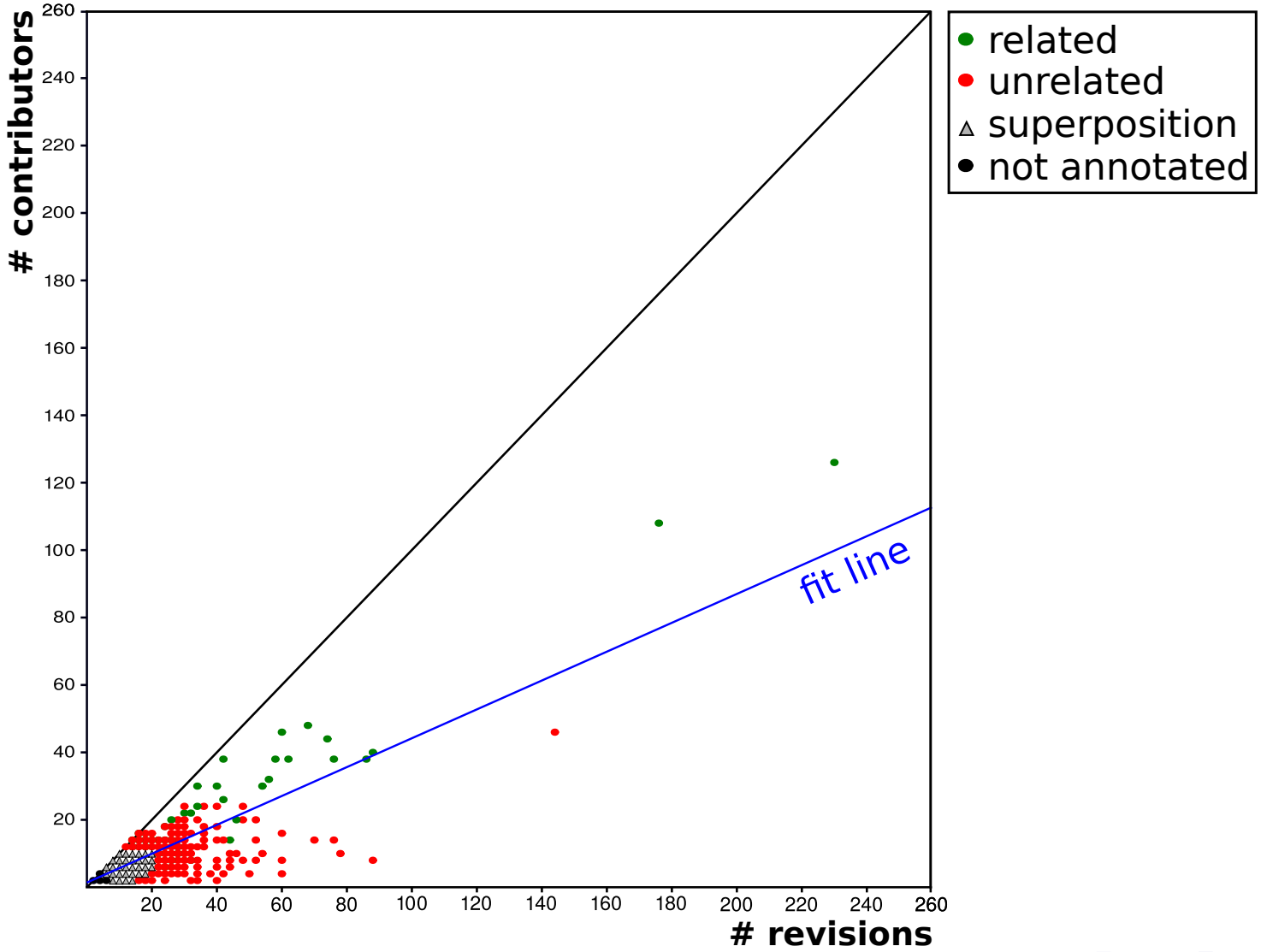
Nouvelles entrées : distributions des annotations



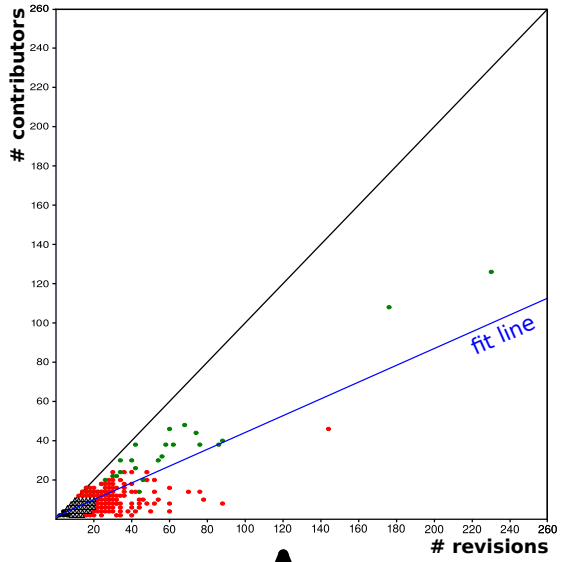
Nouvelles entrées : distributions des annotations

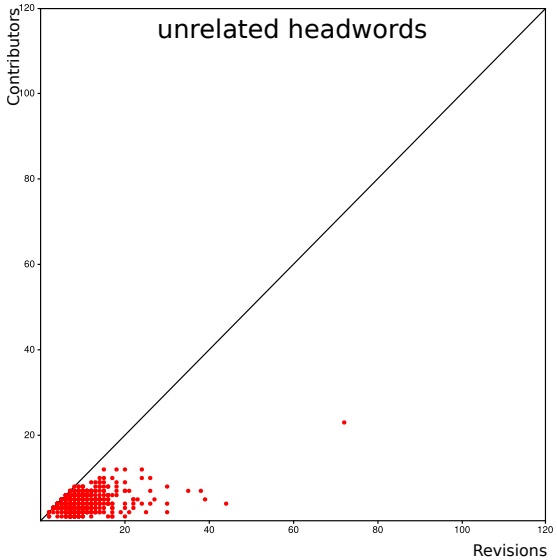


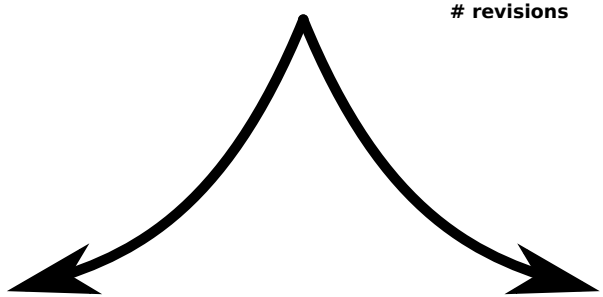
Nouvelles entrées : distributions des annotations

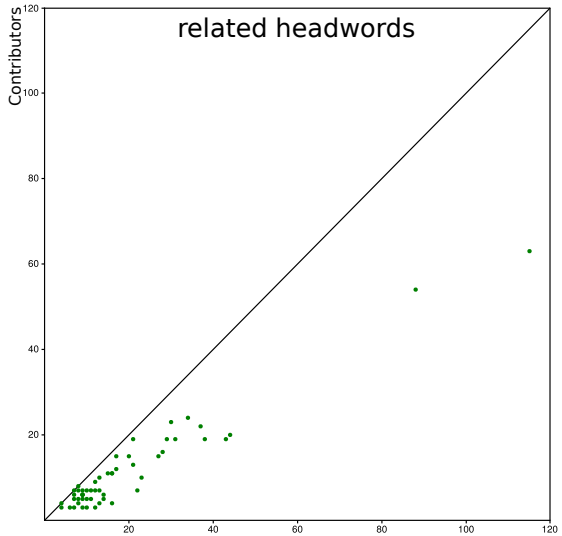


Nouvelles entrées : distributions des annotations

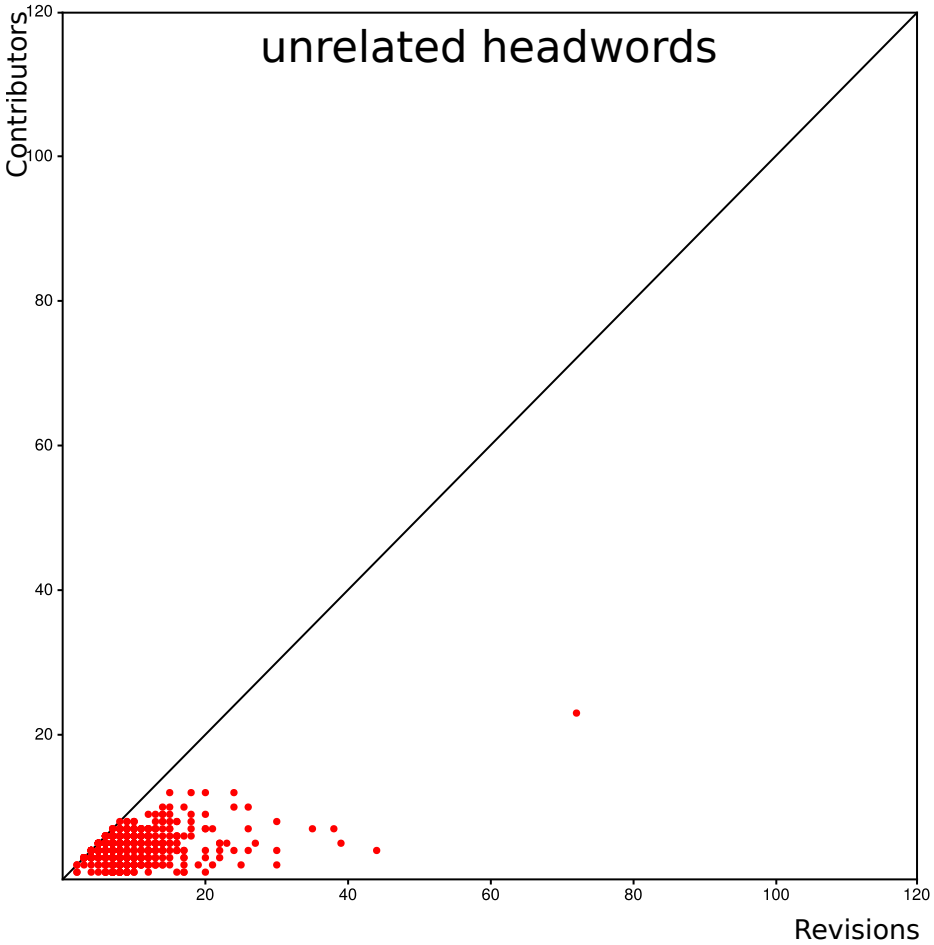


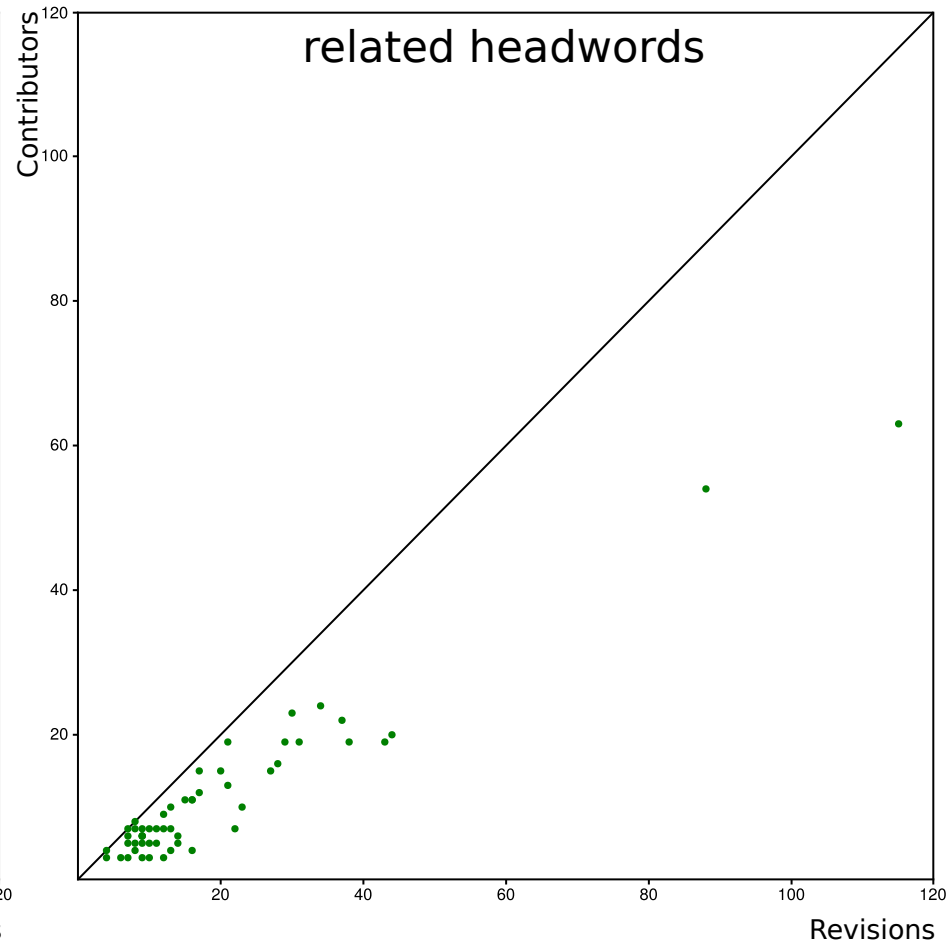




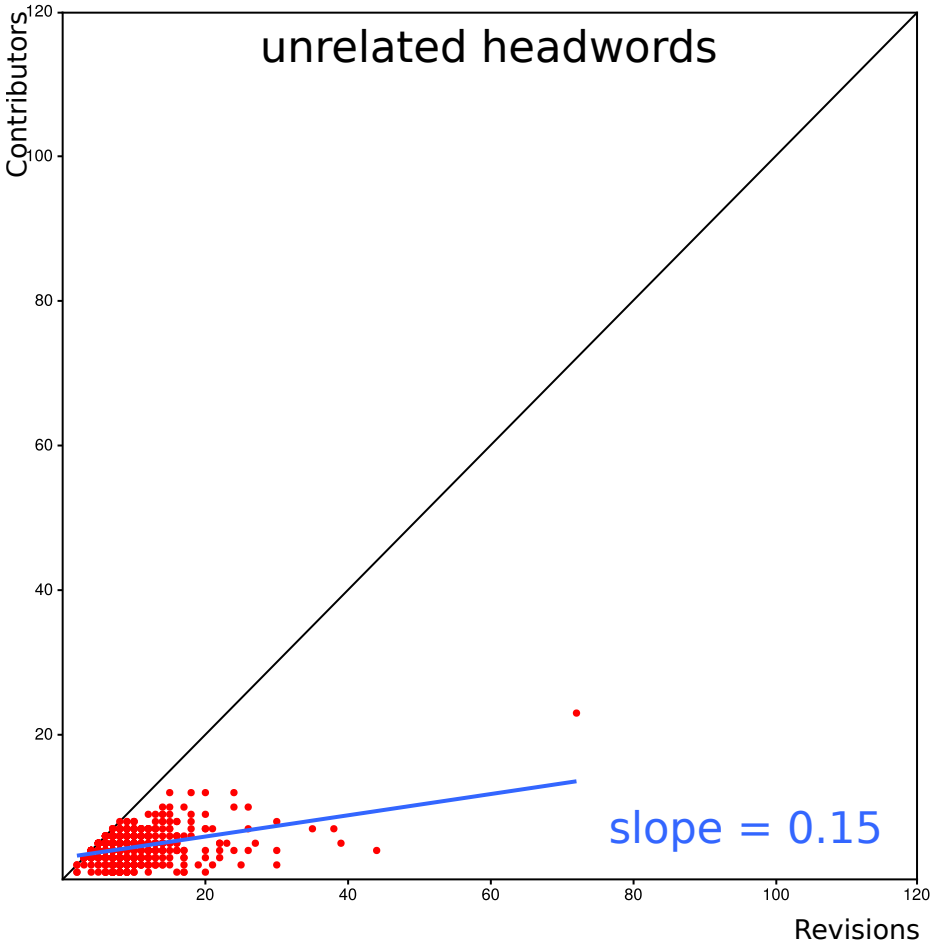


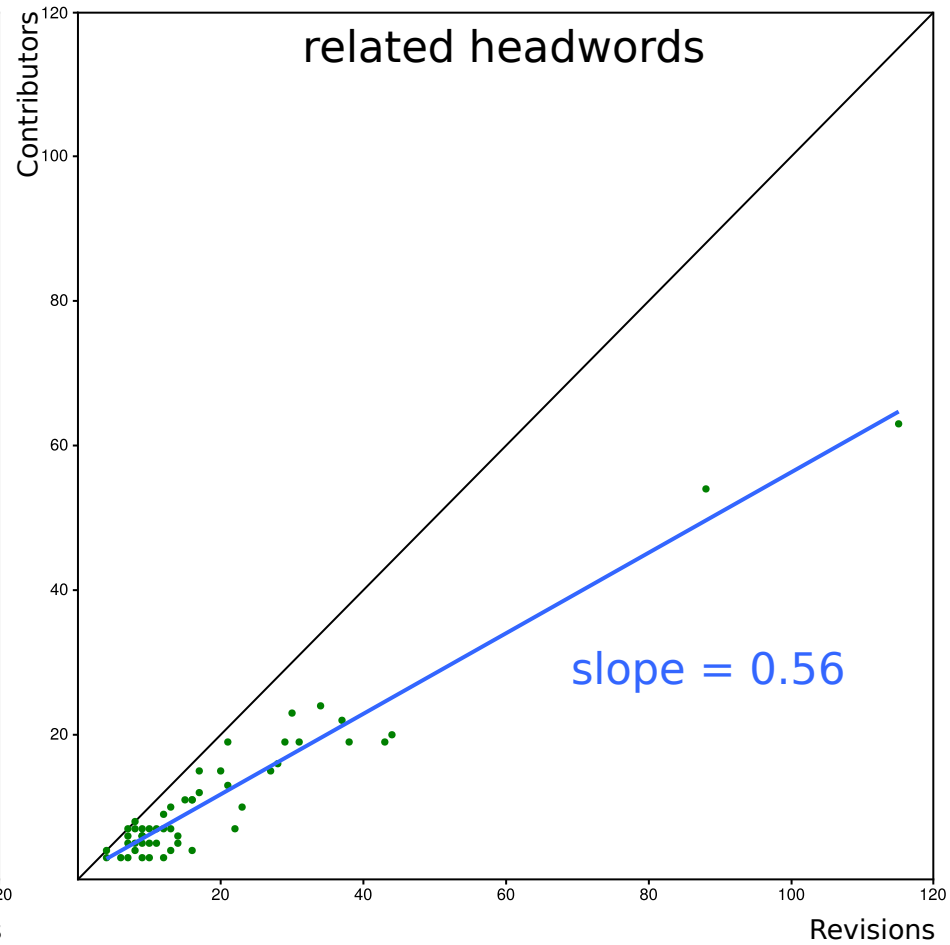
Nouvelles entrées : distributions des annotations



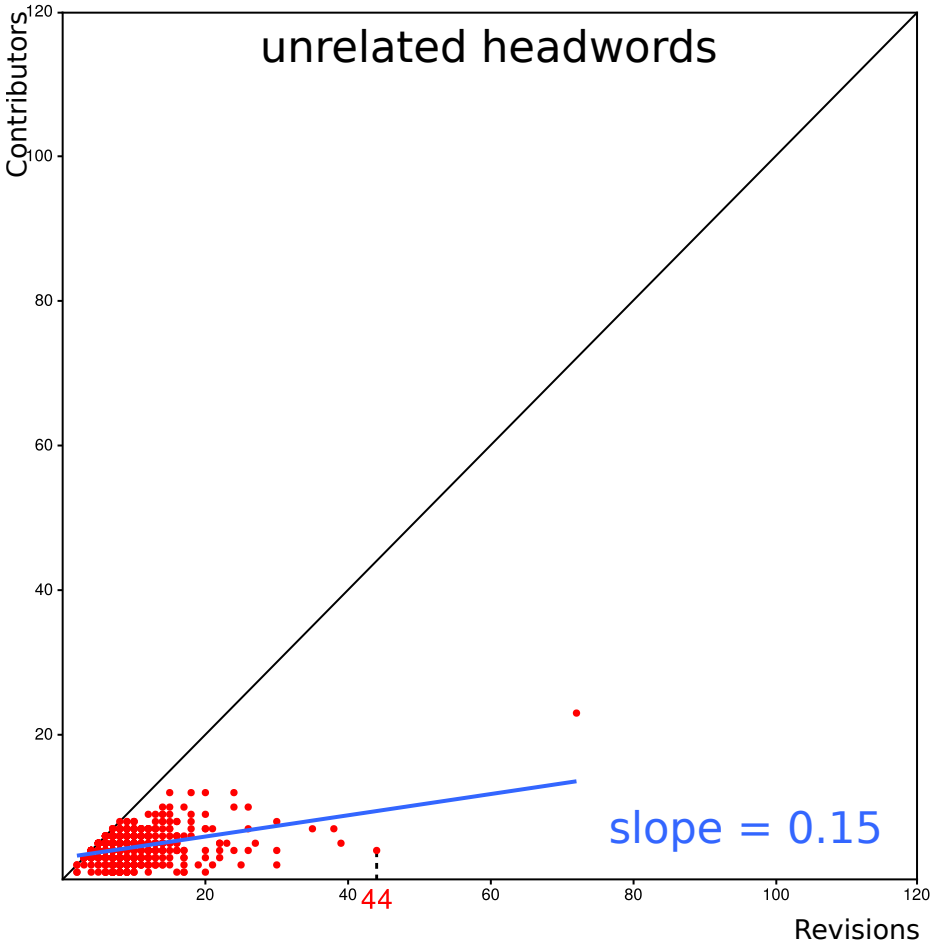


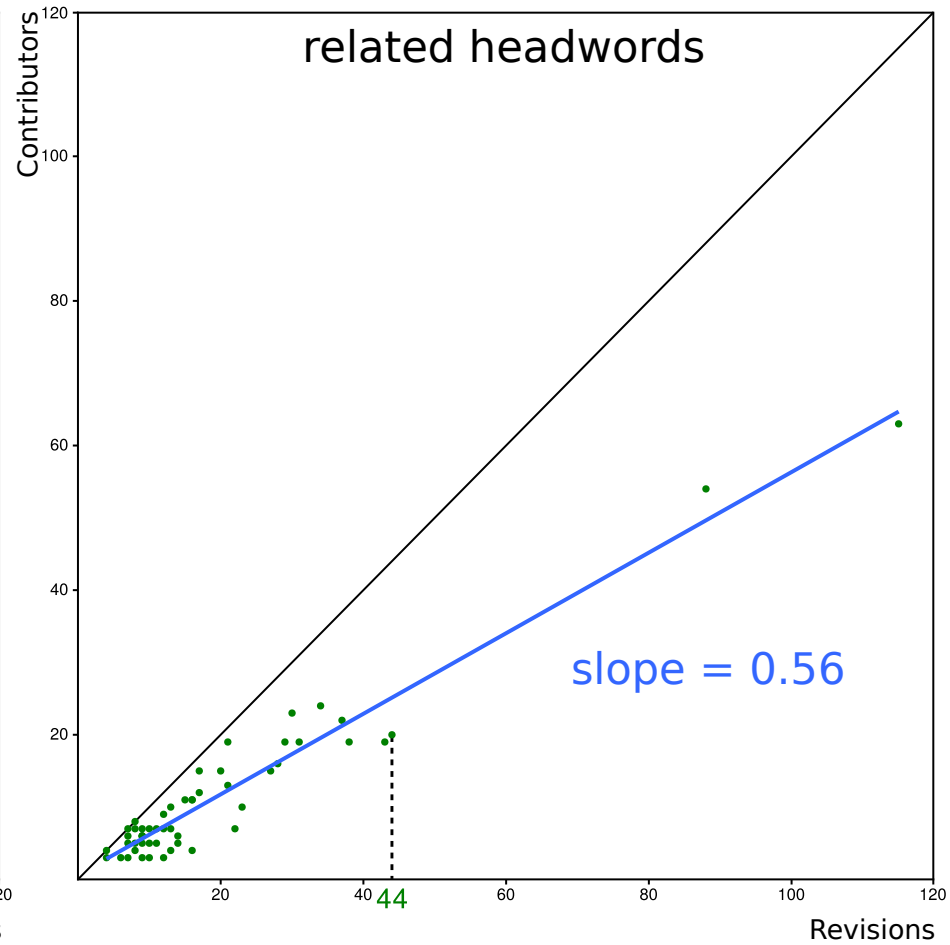
Nouvelles entrées : distributions des annotations



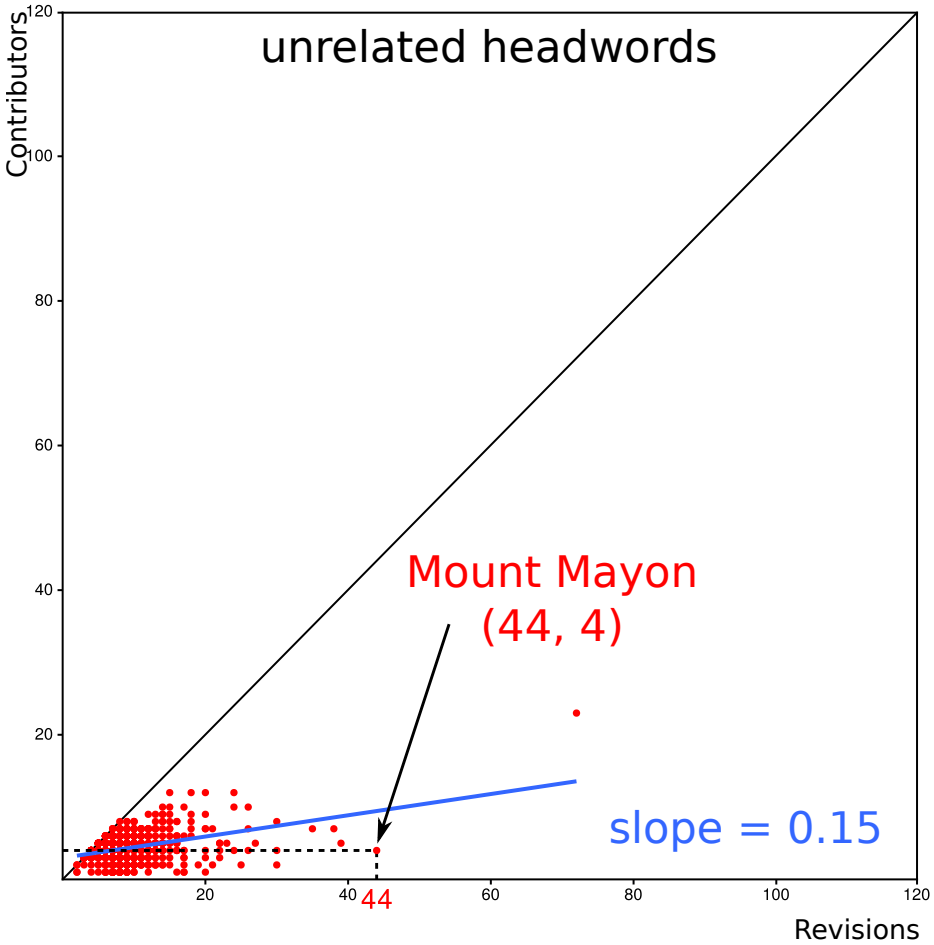


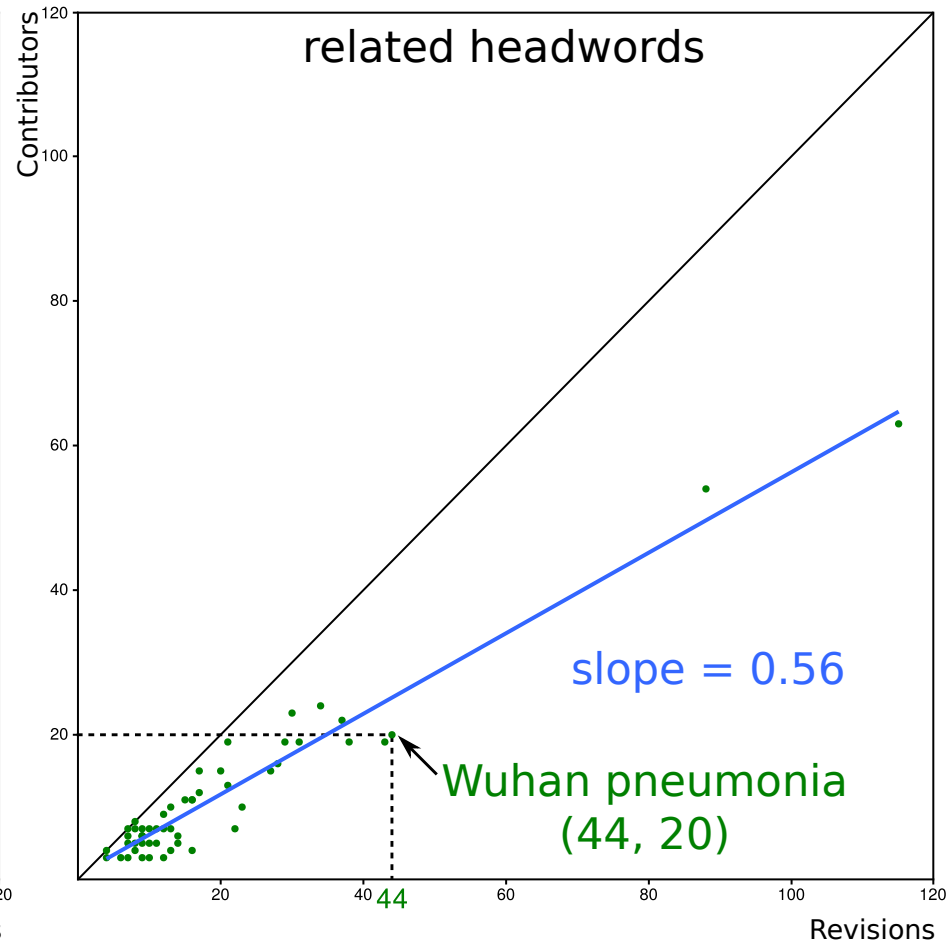
Nouvelles entrées : distributions des annotations





Nouvelles entrées : distributions des annotations





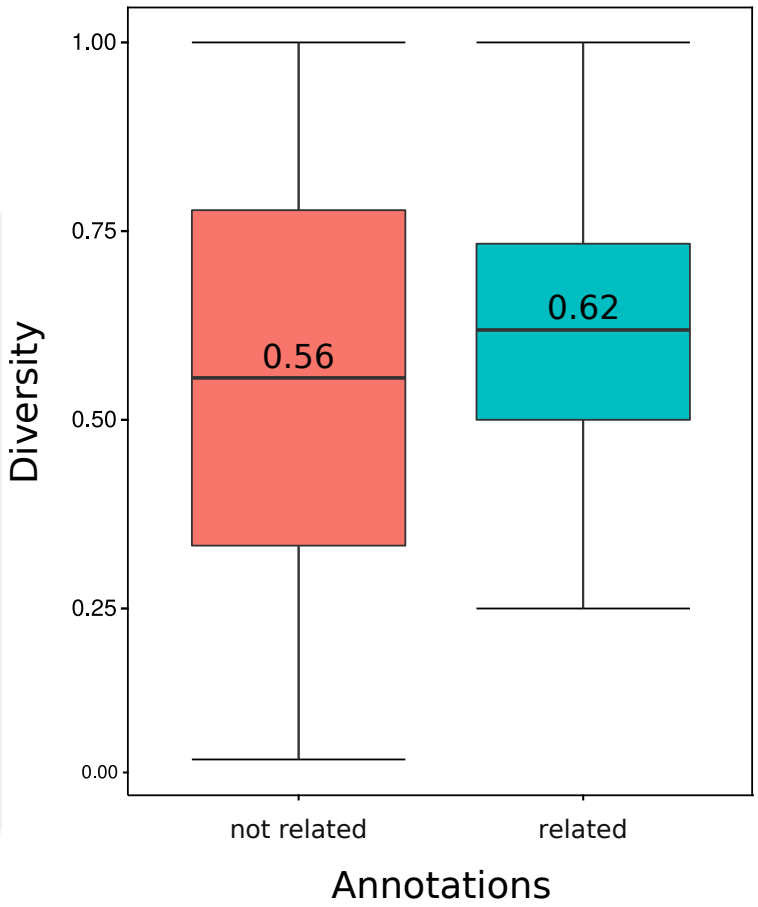
Nouvelles entrées : distributions des annotations

Ma métrique de diversité

Pour une entrée h :

$$diversité(h) = \frac{nbContribs(h)}{nbRevs(h)}$$

- score max = 1 quand $nbContribs(h) = nbRevs(h)$ (*i.e.* chaque revs → 1 contrib ≠)
- score bas quand toutes revs → même contrib., surtout quand $nbRevs(h)$ élevé



(Welch 2-sample t-test : $t(71) = -2.0905$, $p < .05$)

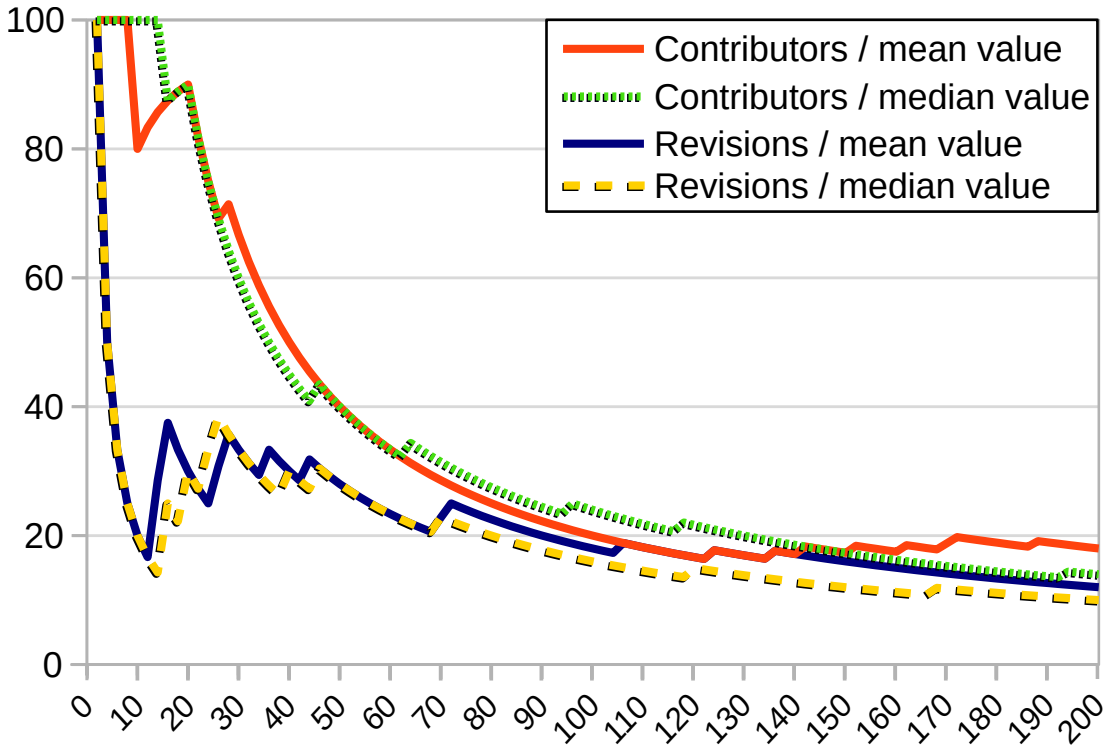
Classement trimestriel des **nouvelles** entrées

Vrais positifs les mieux classés pour chaque trimestre

(graisse normale : mots déjà détectés les trimestres précédents)

- **T1 : social distancing, COVID-19, Wuhan pneumonia, flatten the curve, Wuhan coronavirus, SARS-CoV-2, Kung Flu, Wuhan virus, COVID, social distance, Wuhan flu, SARS-CoV, case fatality rate, Chinese virus, self-isolate, covid, community spread, self-isolation, self-quarantine, noncoronaviral**
- **T2** : COVID-19, social distancing, **infectious disease specialist**, **covidiot**, SARS-CoV-2, flatten the curve, Wuhan virus, self-isolation, COVID, Chinese virus, social distance, **corona virus**, self-quarantine, **elbow shake**, Kung Flu, SARS-CoV, **Rona**
- **T3** : COVID-19, covid, social distancing, **maskne**, **mascne**, **plandemic**, Chinese virus, **contact tracing**, **covid-19 party**, covidiot, **doomscrolling**, Wuhan shake, coronoia, **antimasker**, social distance, SARS-CoV-2
- **T4** : **fever clinic**, **rat-licker**, COVID-19, **before times**, China virus, **Covidtide**, Wuhan flu, **coronasceptic**, **long-hauler**, covidiot, Wuhan virus, **long Covid**, **long covid**

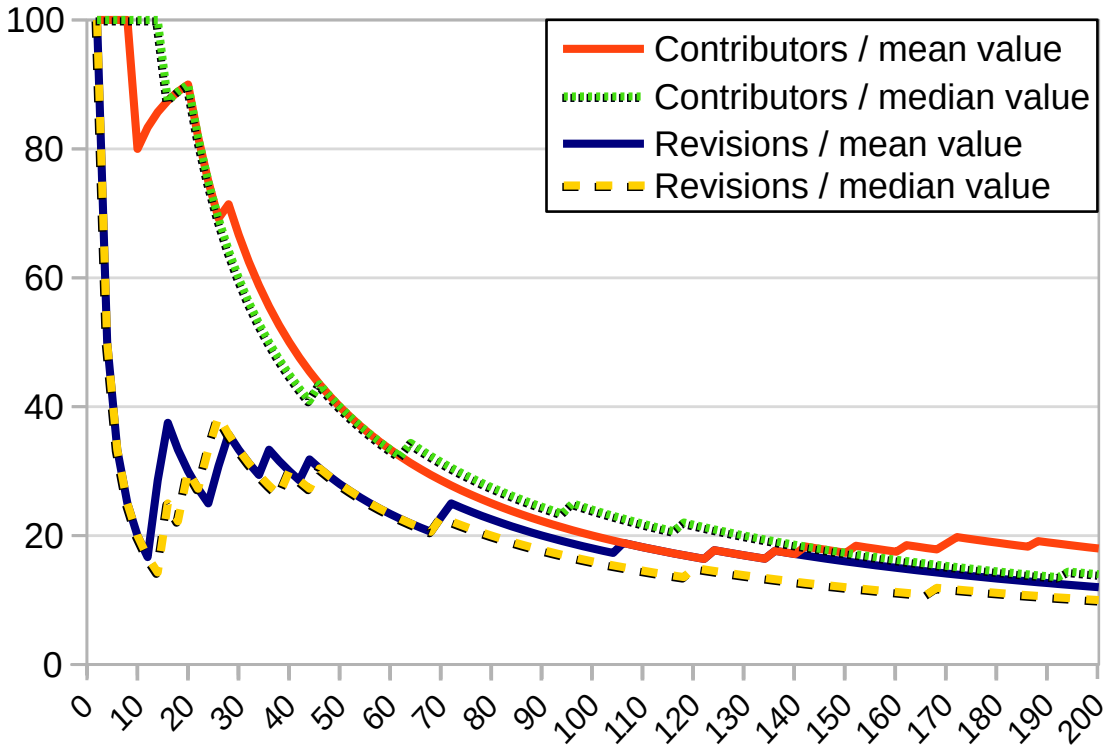
Classement annuel/trimestriel des entrées **existantes**



Classement annuel : vrais positifs les mieux classés

coronavirus, hydroxychloroquine, lockdown, superspreader, surgical mask, pandemic, facemask, corona, herd immunity, MERS, MERS-CoV, Coronavirus, Zoom, ventilator, chloroquine, facial mask, SARS...

Classement annuel/trimestriel des entrées **existantes**



Classement annuel : vrais positifs les mieux classés

coronavirus, hydroxychloroquine, lockdown, superspreader, surgical mask, pandemic, facemask, corona, herd immunity, MERS, MERS-CoV, Coronavirus, Zoom, ventilator, chloroquine, facial mask, SARS...

Classement trimestriel

→ mêmes mots (ordre différent)
+ *panic buying*, *respirator* (T1), *pulmonologist*, *quar* (T2), *virulent* (T4)...

Faux négatifs : à part les jolis exemples qui marchent...

Qu'est-ce qu'on a raté ?

- OED, mise à jour 04/2020 : mots classés en tête de liste des nouvelles entrées de Wiktionary (*e.g. Covid-19*, *social distancing*, *flatten the curve*, *Covid*, *self-isolation*, *contact tracing*, *self-quarantine*, *self-isolate*)
- OED, mise à jour 07/2020 : mots souvent déjà présents dans Wiktionary avant 2020 (*e.g.* en tête de liste : *corona*, *surgical mask*, *MERS*, *Zoom*, *dexamethasone*, *comorbidity*)

Faux négatifs : à part les jolis exemples qui marchent...

Qu'est-ce qu'on a raté ?

- OED, mise à jour 04/2020 : mots classés en tête de liste des nouvelles entrées de Wiktionary (*e.g. Covid-19*, *social distancing*, *flatten the curve*, *Covid*, *self-isolation*, *contact tracing*, *self-quarantine*, *self-isolate*)
- OED, mise à jour 07/2020 : mots souvent déjà présents dans Wiktionary avant 2020 (*e.g.* en tête de liste : *corona*, *surgical mask*, *MERS*, *Zoom*, *dexamethasone*, *comorbidity*)
- mots absents de Wiktionary : *shelter-in-place*, entrée cachée (*run-on entry*) sous *shelter*, non pris en compte dans l'expérience, *community transmission* (*community spread* présent, score élevé)
- mots présents, scores trop faibles (trop peu de révisions) :
  - *Kawasaki's*, défini comme « synonym of Kawasaki disease »
  - *contact tracer* (score élevé pour *contact tracing*)
  - *aéroportage* ajouté (11/2020) directement avec 2 sens, défs et exemples, prononciation, tableau de flexion et synonymes. Une seule révision → indétectable

Faux négatifs : à part les jolis exemples qui marchent...

Tous les mots annotés positivement...

Ont-ils vocation à entrer dans un dictionnaire ?

Faux négatifs : à part les jolis exemples qui marchent...

Tous les mots annotés positivement...

Ont-ils vocation à entrer dans un dictionnaire ?

Ça dépend ! (du dictionnaire, de la ligne éditoriale, etc.)

- mots stigmatisants : *Wuhan flu*, *Chinese virus* → non
  (car éphémères, remplacés rapidement par *Covid-19*)
- créations humoristiques éphémères : *corona belly*, *maskhole* → non
  (ou dictionnaires spécialisés)
- *syndemic*, *doomscrolling* → à considérer
- néologie sémantique : *antimask* → oui !
  Définition actuelle de l'*OED* :
  Now chiefly *historical*.
  A scene (often grotesque or comic) used as a prelude to a masque, or an interlude between its sections, and intended to provide a counterpoint to the main entertainment.

Conclusion : la méthode fonctionne !

Hypothèses confirmées

1. variation importante du nb de révisions → indice de néologie
2. diversité de contributeurs → sujet d'actualité

Détection de la néologie formelle et sémantique

- créations lexicales → articles à ajouter
- mots existants → nouveaux sens à ajouter/articles à réviser

Importance de la *diversité*

étant donné plusieurs articles ayant le même nombre de révisions,
ceux affichant la plus grande diversité de contributeurs
ont le plus de chance d'être liés à des sujets d'actualité

Conclusion : la méthode fonctionne !

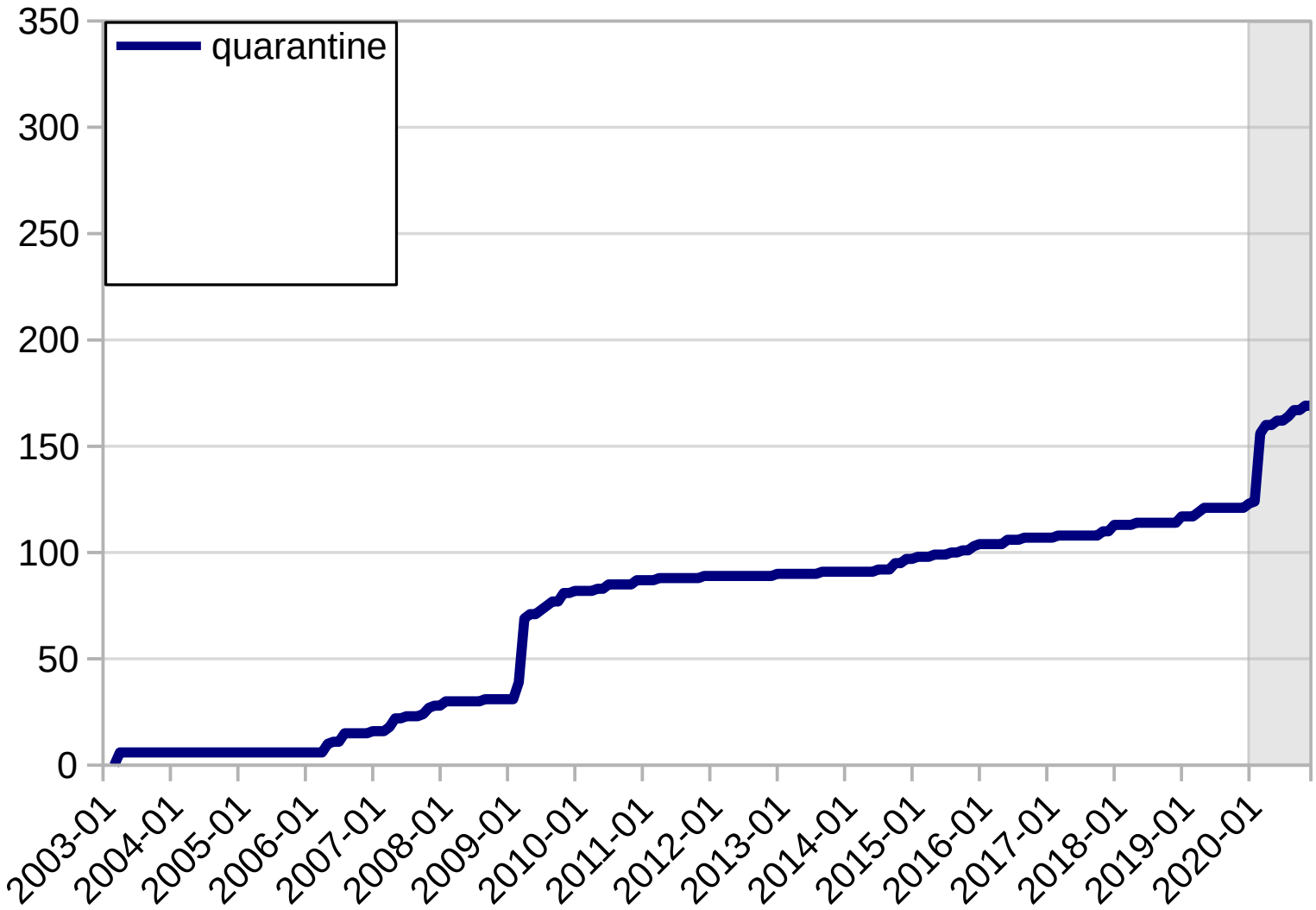
Logs de Wiktionary *vs.* « lexicographie de corpus »

- pas *versus* ! Pas de compétition.
- meilleure option = lexicographie outillée
  + "good old-fashioned lexicography" (Rundell, 2002)
  + logs de Wiktionary + ...

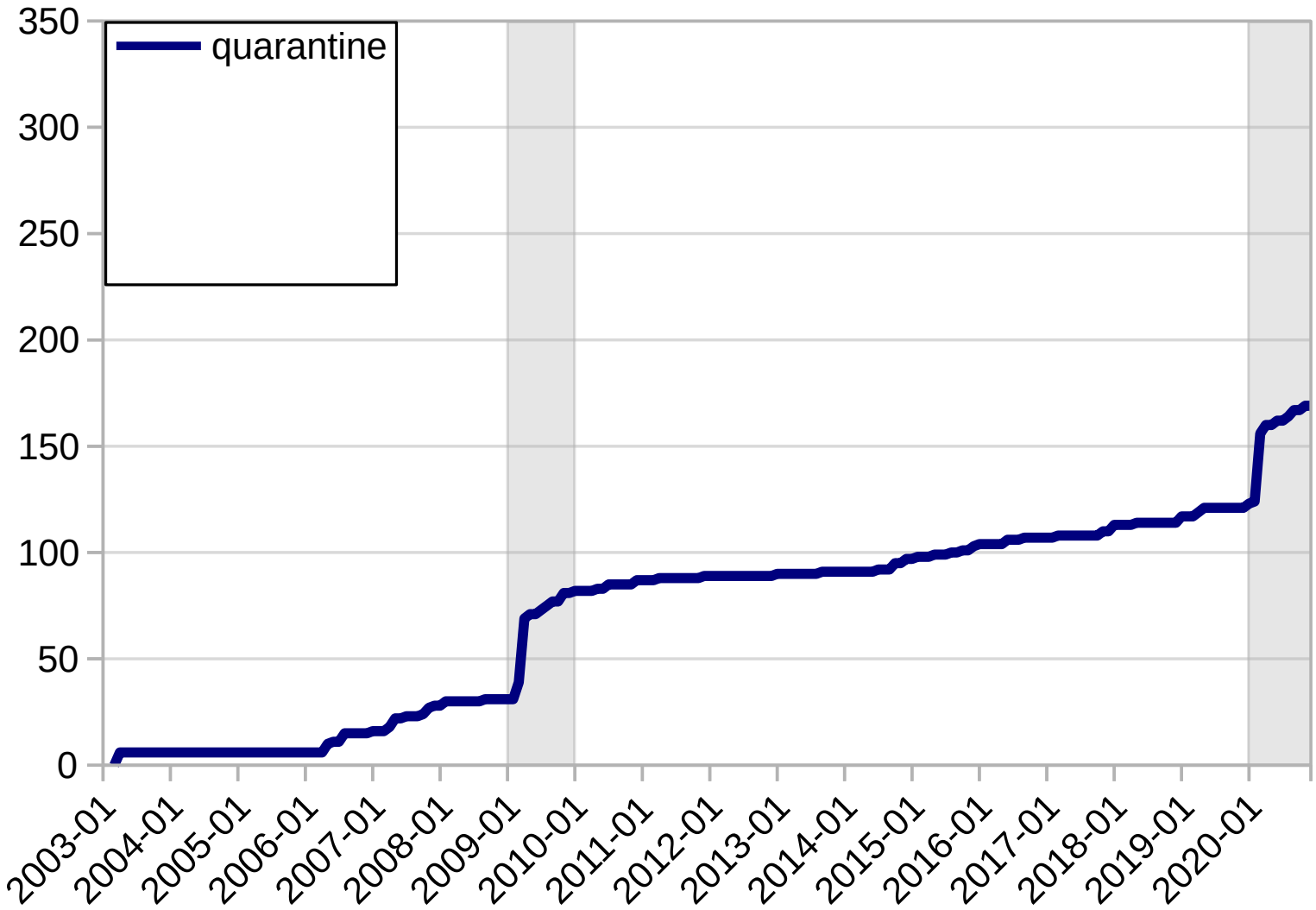
Une méthode réutilisable ?

- pas d'adaptation nécessaire après la première implémentation
- indépendante de la langue...
  → fonctionne probablement le mieux avec les éditions de langue de Wiktionary alimentées par les communautés les plus actives
- indépendante du sujet...
  → pandémie de Covid = une crise mondiale « sans précédent »
  (*unprecedented* = Oxford Languages word of the year 2020)
- à tester sur Wikipédia (taille des données $\neq$!)

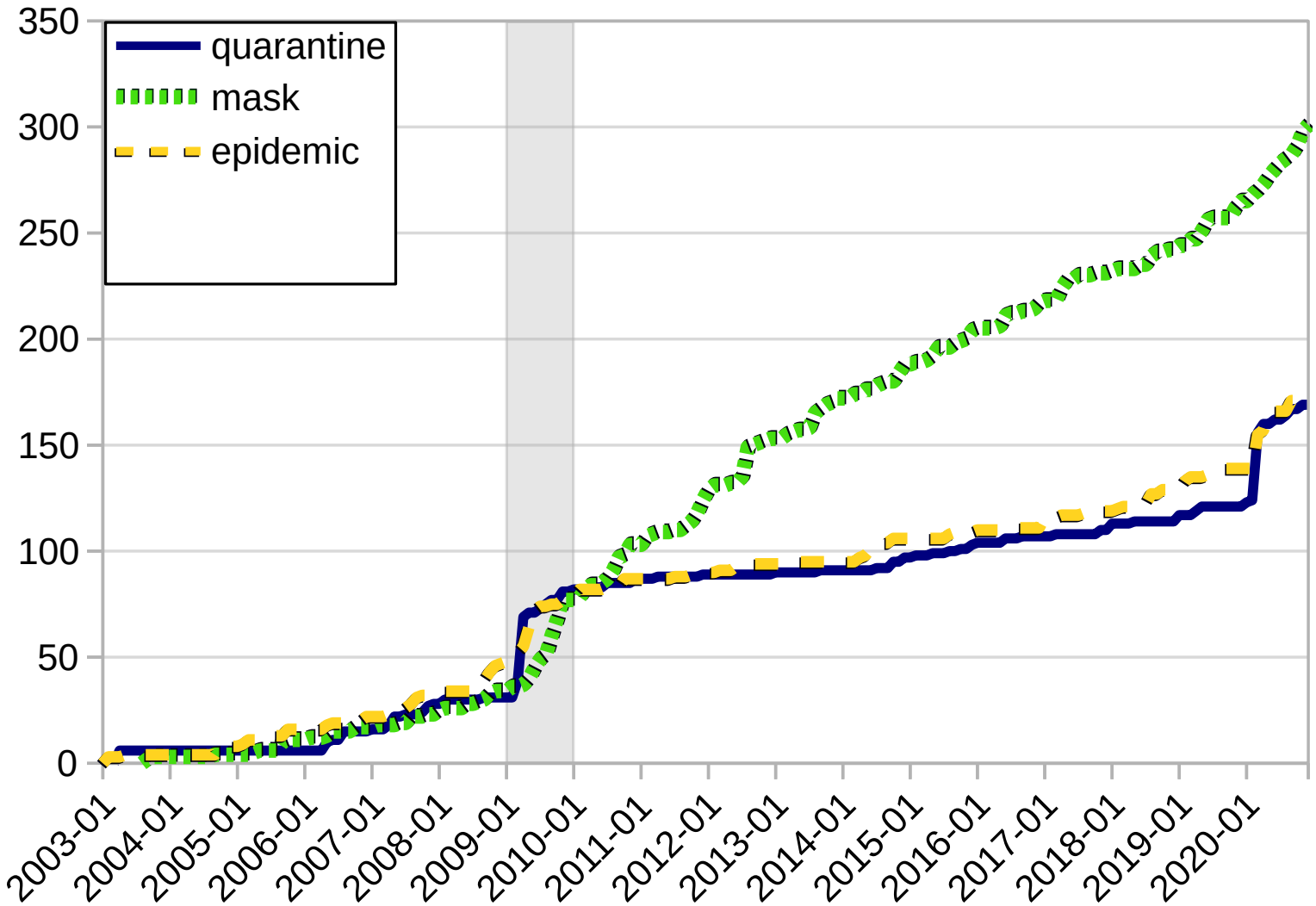
Conclusion : prédire la météo de la veille...



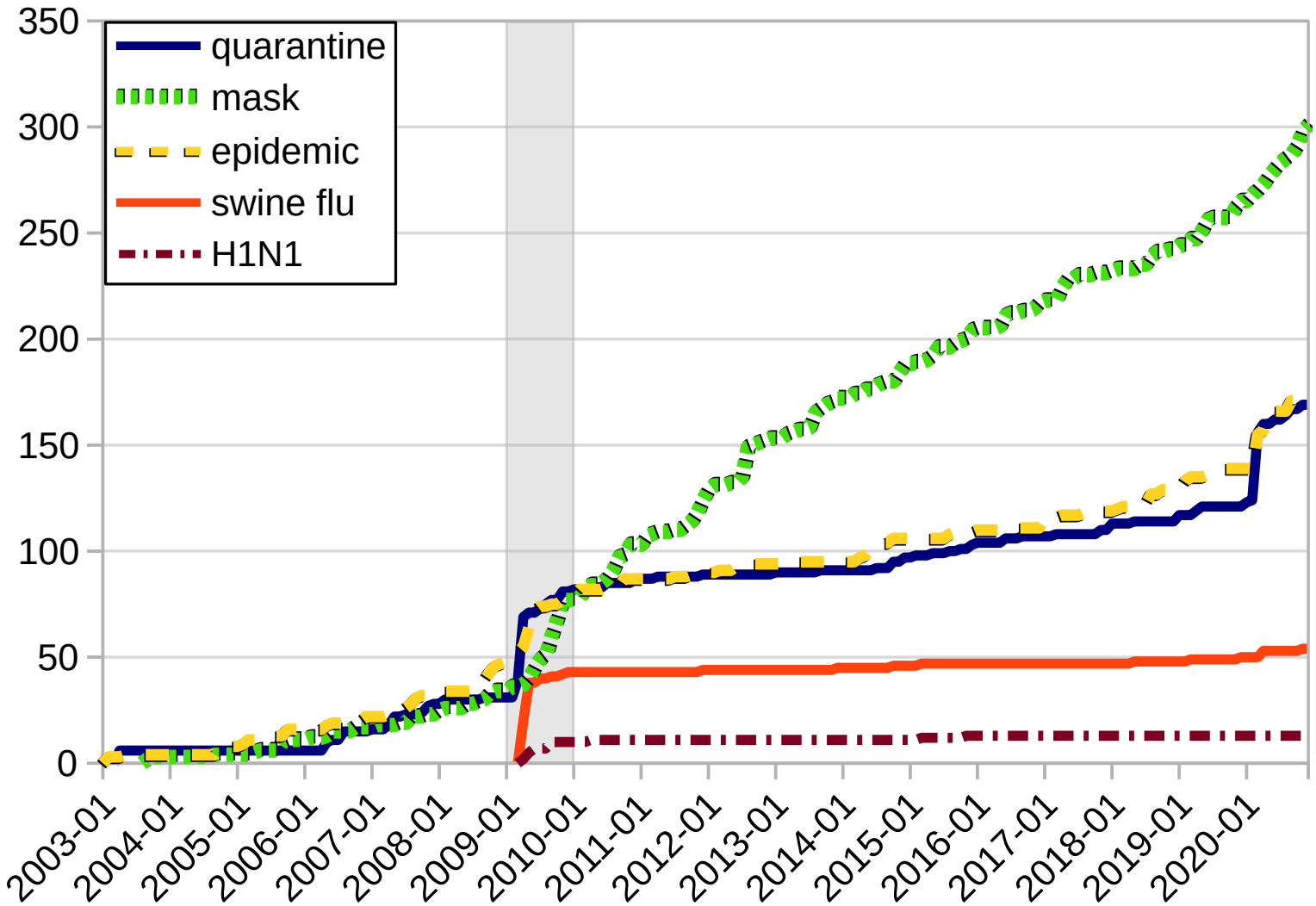
Conclusion : prédire la météo de la veille...



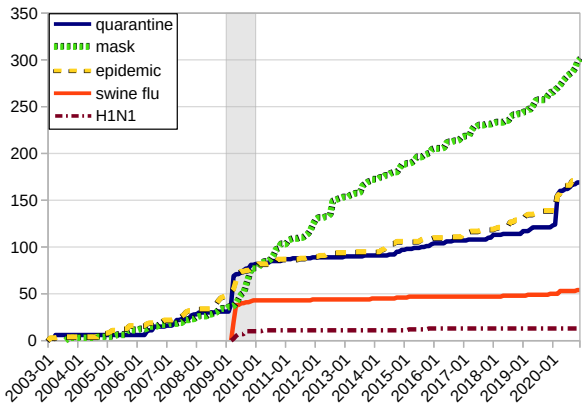
Conclusion : prédire la météo de la veille...



Conclusion : prédire la météo de la veille...



Conclusion : prédire la météo de la veille...



...et du jour

Wiktionnaire, premier semestre 2021

- nouvelles entrées :
  - 4^e: *vaxxie* (« selfie pris pendant l'administration d'un vaccin »)
  - 9^e: *centre de vaccination*
  - 17^e: *Covid long*
  - 63^e: *passeport vaccinal*
- entrées existantes :
  - 17^e: *couvre-feu*
  - 24^e: *vaccinodrome*

Bibliographie I

Boisson, C. (2000). Définitions lexicographiques et partiques sexuelles déviantes. In Béjoint, H. and Thoiron, P., editors, *Le sens en terminologie*, pages 256–279. Presses Universitaires de Lyon.

Čibej, J., Fišer, D., and Kosem, I. (2015). The role of crowdsourcing in lexicography. In *Proceedings of the Elex Conference*, pages 70–83, Herstmonceux Castle, UK.

Cook, P., Lau, J. H., Rundell, M., McCarthy, D., and Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new word-senses. In *Proceedings of the eLex 2013 conference*, pages 49–98, Tallinn, Estonia.

Corbin, P. (1998). La lexicographie française est-elle en panne ? In *Cicle de Conferències 96-97, Lèxic, corpus i diccionaris*, pages 83–112, Barcelona.

Corbin, P. (2008). Quel avenir pour la lexicographie française ? In *Congrès Mondial de Linguistique Française*, pages 1227–1250, Paris, France.

Bibliographie II

Falk, I., Bernhard, D., and Gérard, C. (2014). From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 4337–4344, Reykjavik.

Farina, A. (2005). Lexicographie et discrimination. Corso on line - Introduzione agli studi di genere.

Kilgarriff, A. (1998). The hard parts of lexicography. *International Journal of Lexicography*, 11(1):51–54.

Kilgarriff, A. (2009). Simple maths for keywords. In Mahlberg, M., González-Díaz, V., and Smith, C., editors, *Proceedings of Corpus Linguistics Conference*, Liverpool, UK.

Kosem, I., Gantar, P., and Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowdsourcing. In *Proceedings of Elex 2013*, pages 32–48, Tallin.

Bibliographie III

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico.

Landau, S. (2001). *Dictionaries: The Art and Craft of Lexicography*. Cambridge University Press, Cambridge.

Lih, A. (2004). Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism*, Austin, Texas.

Lorentzen, H. and Trap-Jensen, L. (2016). What, When and How? - the Art of Updating an Online Dictionary. In Tinatin Margalitadze, G. M., editor, *Proceedings of the 17th EURALEX International Congress*, pages 138–145, Tbilisi, Georgia. Ivane Javakhishvili Tbilisi University Press.

Mortureux, M.-F. (2011). La néologie lexicale : de l'impasse à l'ouverture. *Langages*, 183(2):11–24.

Bibliographie IV

Pruvost, J. (2006). *Les dictionnaires français, outils d'une langue et d'une culture*. Ophrys, Paris.

Renouf, A. (2013). A Finer Definition of Neology in English: the life-cycle of a word. In Hasselgård, H., Ebeling, J., and Ebeling, S. O., editors, *Corpus Perspectives on Patterns of Lexis*, pages 177–208. John Benjamins.

Rey, A. (1995). Du discours au discours par l'usage : pour une problématique de l'exemple. *Langue française*, 106:95–120.

Rundell, M. (2002). Good Old-fashioned Lexicography: Human Judgment and the Limits of Automation. In Corréard, M.-H., editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, pages 138–155. EURALEX, Grenoble.

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

Bibliographie V

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

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

Sablayrolles, J.-F. (2008). Néologie et dictionnaire(s) comme corpus d'exclusion. In Sablayrolles, J.-F., editor, *Néologie et terminologie dans les dictionnaires*, pages 19–36. Champion, Paris.

Sajous, F., Josselin-Leray, A., and Hathout, N. (2018). The Complementarity of Crowdsourced Dictionaries and Professional Dictionaries viewed through the Filter of Neology. *Lexis*, 12.

Sajous, F. and Martinez, C. (2021). Metalexicographical Investigations with the DiCo Database. *International Journal of Lexicography*, 34(4).

Bibliographie VI

Sommant, M. (2000). Innovation lexicale : sources des néologismes, normalisation et intégration dans les nomenclatures des dictionnaires de langue français. In Béjoint, H. and Thoiron, P., editors, *Le Sens en terminologie*, pages 247–260. Presses Universitaires de Lyon.

Wolfer, S. and 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.