La biodiversité et le désaccord scientifique

ARMB, 2024-03-16

Charles H. Pence
@pence@scholar.social

UCLouvain

ISP Institut supérieur de philosophie

LIBST Louvain Institute of Biomolecular Science and Technology

Académie Royale de Belgique 1772

Plan

1. Ambiguïté et désaccord dans la biodiversité et la taxonomie
2. Analyse empirique : corpus taxonomique
3. Désaccord sur quoi ?

Le message d'ensemble : Il y a un sentiment fort dans la biologie et la philosophie que le désaccord pose un problème sérieux pour la conservation. Testons-le !

La biodiversité et la taxonomie

Un équilibre

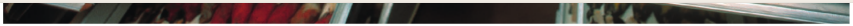
Le concept de la biodiversité doit être :

- **Plus grand que des espèces isolées (charismatiques), afin de capter les relations écologiques**
- **Plus petite que « tout le vivant », afin de pouvoir la protéger**

Biodiversité et taxinomie

Mesure la plus commune : richesse spécifique

Mais tout étude de la biodiversité qui dépend des espèces va hériter tout le désaccord de la taxinomie !



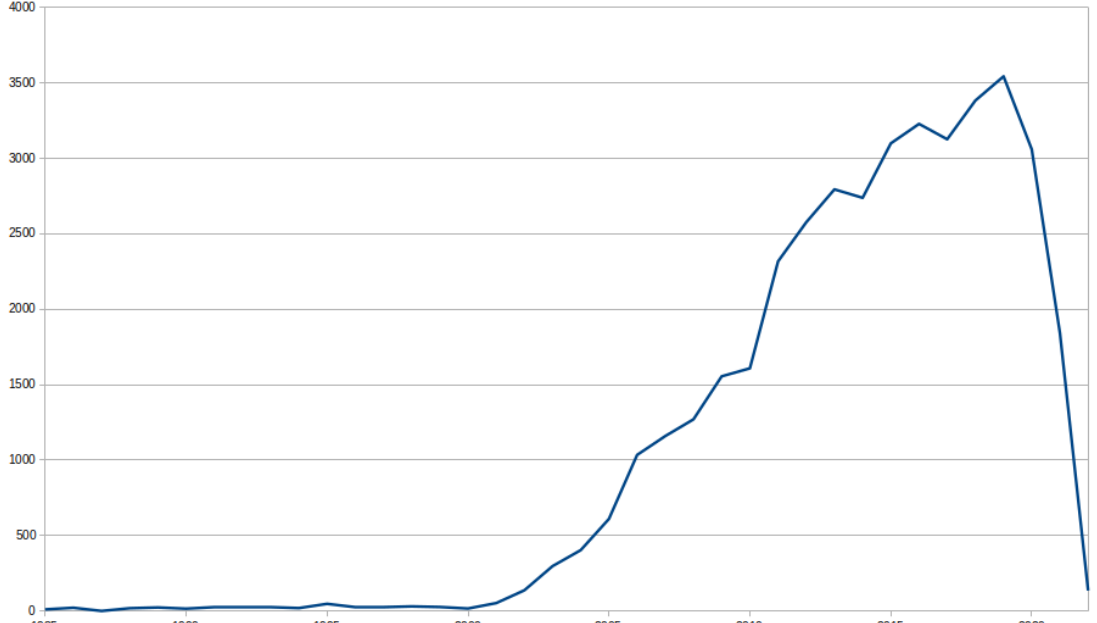
Part of the vast ornithology collection at the American Museum of Natural History.

Taxonomy anarchy hampers conservation

The classification of complex organisms is in chaos.
Stephen T. Garnett and **Les Christidis** propose a solution.

Corpus taxonomique

Révue	Maison d'édition	Taille
Zootaxa	Magnolia Press	31,348
ZooKeys	Pensoft	4,940
PhytoKeys	Pensoft	820
Journal of Hymenoptera Research	Pensoft	382
MycoKeys	Pensoft	315
Zoosystematics and Evolution	Pensoft	153
Insecta Mundi	Center for Systematic Entomology	1,367
European Journal of Taxonomy	Museum National d'Histoire Naturelle	1,105


4000
3500
3000
2500
2000
1500
1000
500
0
1985
1990
1995
2000
2005
2010
2015
2020

Où se trouve-t-il le désaccord ?

Comment détecter le désaccord ?

Lecture attentive de plusieurs articles où l'on sait que le désaccord taxonomique a lieu

Comment détecter le désaccord ?

Exemple : la liste *disagreement*

- critique
- doubt
- opinion
- disagree
- redundant
- reject
- rebuttal
- debate
- invalid
- misunderstanding
- misconception
- allegation
- allegedly
- mistake
- obsolete
- error
- misclassify
- erroneous
- contentious

Topic Modeling

Bref : une sorte de réduction de la dimensionnalité non supervisé qu'on peut réaliser sur un corpus de texte. Ça transforme les documents, pris comme vecteurs dans un espace de 172M dimensions, en vecteurs 125-D.

Où se trouve le désaccord ?

Demander au model : quels *topics* donnent plus de probabilité aux mots dans nos listes de termes qui signalent le désaccord ?

Où se trouve le désaccord ?

Demander au model : quels *topics* donnent plus de probabilité aux mots dans nos listes de termes qui signalent le désaccord ?

- ***disagreement*** **: Topic 43**
- ***pejorative terms*** **: Topics 43 et 120**

Topic 43 (disagreement, pejorative)

- 0.015*“specie”
- 0.011*“name”
- 0.010*“description”
- 0.010*“new”
- 0.008*“publish”
- 0.007*“author”
- 0.007*“nomenclature”
- 0.007*“code”
- 0.007*“publication”
- 0.006*“type”
- 0.006*“article”
- 0.006*“zoological”
- 0.006*“original”
- 0.006*“synonym”
- 0.006*“work”
- 0.006*“list”
- 0.006*“valid”
- 0.005*“international”
- 0.005*“available”
- 0.005*“note”

Les mots qu’on utilise pour présenter une nouvelle espèce ainsi que discuter si une espèce est un synonyme

Topic 120 (pejorative)

- 0.018*“character”
- 0.013*“genera”
- 0.011*“taxon”
- 0.011*“group”
- 0.010*“specie”
- 0.010*“genus”
- 0.009*“phylogenetic”
- 0.008*“include”
- 0.007*“analysis”
- 0.007*“family”
- 0.007*“relationship”
- 0.005*“phylogeny”
- 0.005*“clade”
- 0.005*“morphological”
- 0.005*“classification”
- 0.005*“support”
- 0.005*“press”
- 0.005*“new”
- 0.005*“consider”
- 0.004*“present”

Les mots qu’on utilise pour disputer le rang d’un groupe

Plus de précision ?

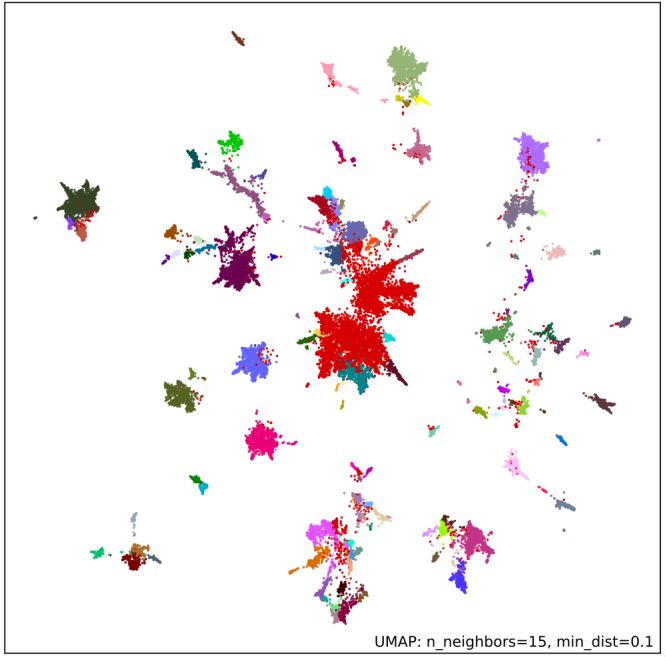
Un peu plus de précision est souhaitable dans ces analyses, pour distinguer entre plusieurs sens des mots dans ces *topics* – par exemple, entre décrire une nouvelle espèce et supprimer une espèce

Document Vector Model

Entraîner un modèle qui représente les mots dans notre corpus directement par des vecteurs dans un espace 100-dimensionnel,[1] et puis représenter chaque document comme un vecteur dans ce même espace.[2]

[1] Techniquement : un modèle Word2Vec entraîné par *hierarchical softmax*

[2] Techniquement : un modèle Doc2Vec, qui infère les représentations vectorielles des documents par échantillonnage d'une fenêtre glissante des mots


UMAP: n_neighbors=15, min_dist=0.1

Désaccord sur quoi ? Taxa

De quels groupes êtes-vous plus probable de discuter dans les articles dans la zone *disagreement* de notre espace vectoriel ? Extraire tous les noms d'espèces[3] des 5 000 documents les plus proches aux mots dans notre liste *disagreement*, ainsi que les 5 000 documents les plus loins, et comparer le risque relatif.

[3] Techniquement : en utilisant le logiciel impressionnant `gnfinder`

Désaccord par taxon

Plus de désaccord :

Mammifères (≈ 4), Oiseaux (3), Champignons (3), Poissons (2)

Moins de désaccord :

Insectes (≈ 0.5)

Désaccord sur quoi ? Méthodologie

Topic 64 : phylogénétique moléculaire

- 0.021*“specie”
- 0.017*“sequence”
- 0.016*“analysis”
- 0.011*“molecular”
- 0.010*“dna”
- 0.008*“phylogenetic”
- 0.007*“tree”
- 0.007*“clade”
- 0.007*“gene”
- 0.007*“specimen”
- 0.007*“study”
- 0.007*“morphological”
- 0.006*“support”
- 0.006*“group”
- 0.006*“genetic”
- 0.006*“coi”
- 0.006*“datum”
- 0.006*“base”
- 0.005*“table”
- 0.005*“population”

Parmi les 20 *topics* les plus probables dans les reptiles, les amphibiens, les oiseaux, les poissons, les champignons, et les mammifères ; top-5 % dans tout autre groupe

Désaccord sur quoi ?

Hormis les mots dans notre liste *disagreement*, quels mots distinguent les articles « désaccord » des articles « non-désaccord » ?[4]

[4] Techniquement : utiliser l'algorithme Craig Zeta pour séparer les top-5 000 documents des bas-5 000 documents

Désaccord sur quoi ?

Désaccord :

- appear
- note
- consider
- north
- revision
- probably
- lectotype
- list
- suggest
- range
- synonym
- case
- non
- see
- early
- synonymy
- western
- available
- european
- population

Non-désaccord :

- china
- online
- issn
- copyright
- print
- male
- figs
- edition
- holotype
- introduction
- nov
- new
- margin
- lateral
- accept
- dorsal
- eye
- deposit
- length
- head

Recherche future

- Analyser le désaccord seulement dans la section méthodologique des articles
- Construire un « haut désaccord » sous-corpus pour tenter de distinguer le désaccord conceptuel du débat à long terme
- Ajouter le *geocoding* afin d'examiner le désaccord par rapport aux endroits étudiés ?

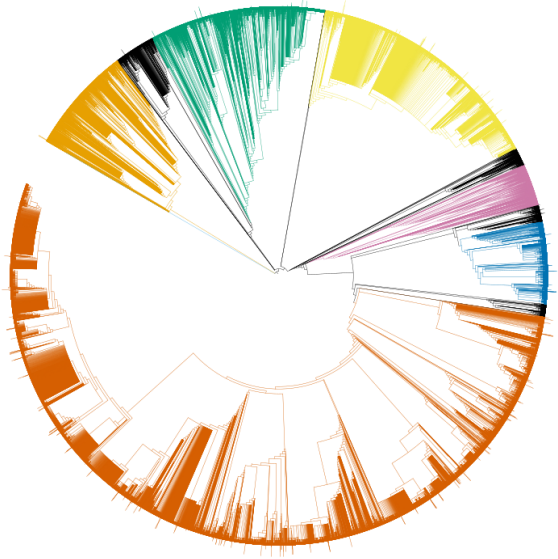
Merci à Stijn Conix !

Questions ?

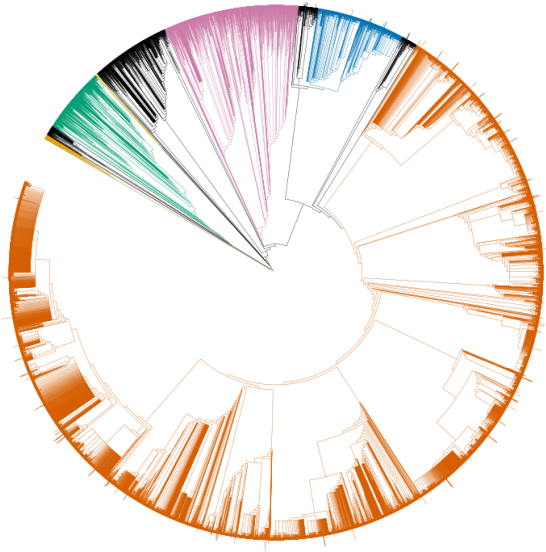
charles@charlespence.net
https://pencelab.be
@pence@scholar.social

PENCE LAB

fnrs
LA LIBERTÉ DE CHERCHER



Open Tree of Life



Corpus

Topic 91 (epistemic value)

- 0.038*“setae”
- 0.022*“margin”
- 0.021*“article”
- 0.019*“long”
- 0.017*“length”
- 0.013*“pereopod”
- 0.010*“fig”
- 0.010*“seta”
- 0.010*“simple”
- 0.009*“propodus”
- 0.009*“short”
- 0.009*“male”
- 0.008*“basis”
- 0.008*“female”
- 0.008*“specie”
- 0.008*“inner”
- 0.008*“robust”
- 0.007*“distal”
- 0.007*“uropod”
- 0.007*“outer”

…decapod crustaceans ? 🤔