Conceptual Cartography and Textual Analysis

Methods in Philosophy of Science, 2023-05-23

Charles H. Pence

@pence@scholar.social

UCLouvain
Institut supérieur de philosophie (ISP)

ACADÉMIE ROYALE DE BELGIQUE
1772

Outline

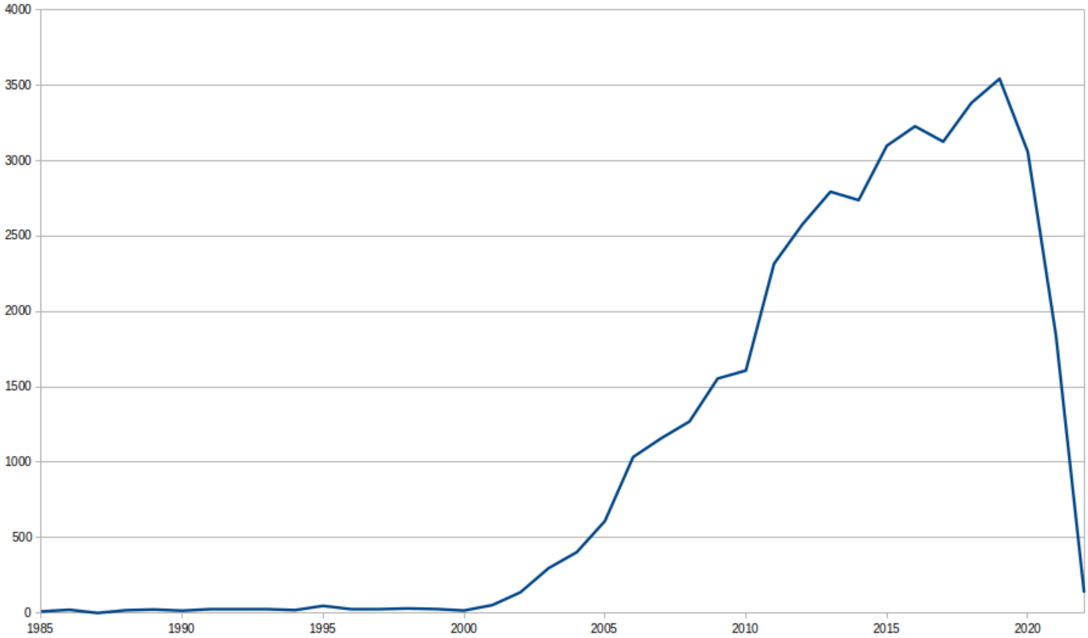
1. Empirical analyses: disagreement in taxonomy and biodiversity
   1.1 Corpus construction
   1.2 Topic modeling
   1.3 Document vectors and stylometry
   1.4 Future ideas
2. Some extremely unstructured thoughts on the distinction between analysis and cartography

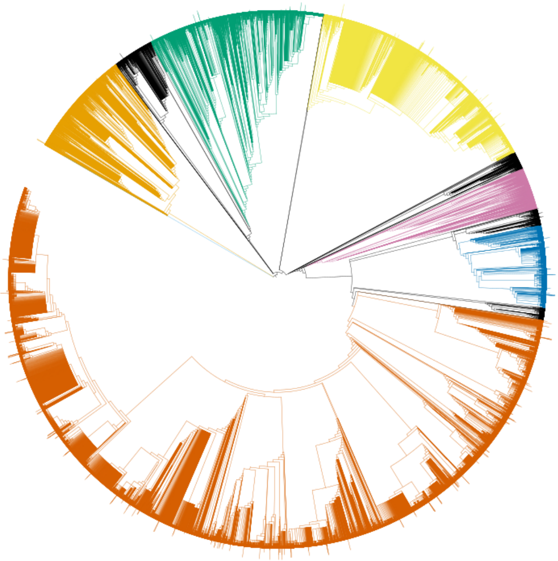
The take-home (question?): How should we understand the nature and role of a potential "conceptual cartography"?

Biodiversity and Taxonomy

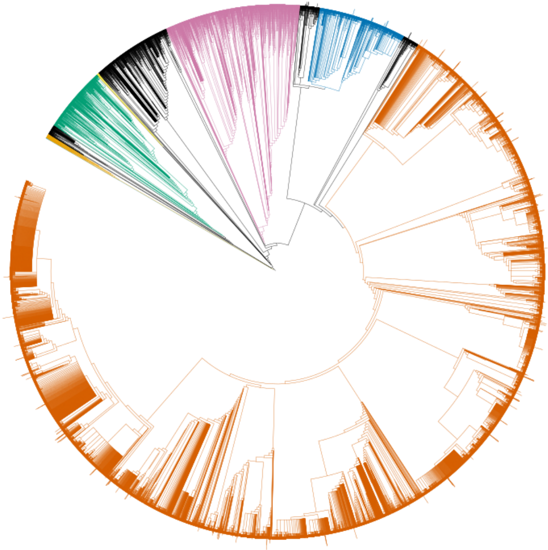
Empirical Tools

Journal	Publisher	Size
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

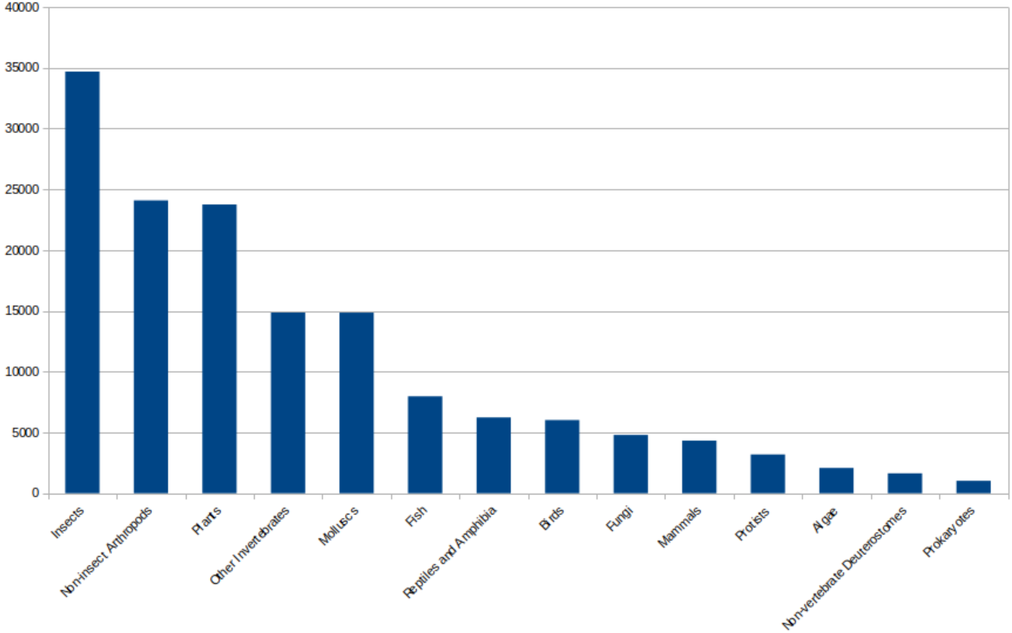

4000
3500
3000
2500
2000
1500
1000
500
0
1985
1990
1995
2000
2005
2010
2015
2020



Complete Open Tree of Life



Corpus


40000
35000
30000
25000
20000
15000
10000
5000
0
Insects
Non-insect Arthropods
Plants
Other Invertebrates
Molluscs
Fish
Reptiles and Amphibia
Birds
Fungi
Mammals
Protists
Algae
Non-vertebrate Deuterostomes
Prokaryotes

Topic Modeling

Briefly: a kind of unsupervised dimensionality reduction that you can run on a corpus of text. Take documents, normally locations in a 172M-dimensional space (number of word types), and reduce that to 125-D.

Interpreting a Topic

Topic 16: popular in mammals

- 0.027*”colombia”
- 0.016*”specie”
- 0.013*”type”
- 0.013*”peru”
- 0.010*”locality”
- 0.010*”venezuela”
- 0.010*”ecuador”
- 0.009*”panama”
- 0.008*”distribution”
- 0.007*”brazil”
- 0.007*”key”
- 0.006*”rica”
- 0.006*”del”
- 0.006*”costa”
- 0.006*”genus”
- 0.006*”male”
- 0.006*”america”
- 0.006*”san”
- 0.006*”neotropical”
- 0.005*”cat”

Interpreting a Topic

Topic 16: popular in mammals

- 0.027*”colombia”
- 0.016*”specie”
- 0.013*”type”
- 0.013*”peru”
- 0.010*”locality”
- 0.010*”venezuela”
- 0.010*”ecuador”
- 0.009*”panama”
- 0.008*”distribution”
- 0.007*”brazil”
- 0.007*”key”
- 0.006*”rica”
- 0.006*”del”
- 0.006*”costa”
- 0.006*”genus”
- 0.006*”male”
- 0.006*”america”
- 0.006*”san”
- 0.006*”neotropical”
- 0.005*”cat”

Okay: Central and South American collection sites

Topic 31:

- 0.016*”male”
- 0.016*”genitalia”
- 0.013*”specie”
- 0.009*”female”
- 0.009*”fig”
- 0.008*”brown”
- 0.008*”lepidoptera”
- 0.007*”scale”
- 0.007*”long”
- 0.006*”slide”
- 0.006*”white”
- 0.006*”line”
- 0.006*”new”
- 0.006*”bursae”
- 0.006*”short”
- 0.005*”dark”
- 0.005*”coll”
- 0.005*”forewing”
- 0.005*”holotype”
- 0.005*”leg”

Cautious hypothesis: Lepidopteran anatomy, especially reproductive

Interpreting a Topic

But wait.

Our lepidopteran reproductive anatomy topic is unusually significant in one group... **in papers that mention molluscs.**

Interpreting a Topic

But wait.

Our lepidopteran reproductive anatomy topic is unusually significant in one group... **in papers that mention molluscs.**

...too many bursas!

Some Cool Topics

Topic 9: traditional specimen collection terms

- 0.029*”specie”
- 0.012*”forest”
- 0.012*”habitat”
- 0.010*”area”
- 0.008*”find”
- 0.007*”collect”
- 0.007*”site”
- 0.007*”study”
- 0.007*”record”
- 0.006*”population”
- 0.006*”range”
- 0.006*”high”
- 0.005*”specimen”
- 0.005*”occur”
- 0.005*”know”
- 0.004*”individual”
- 0.004*”region”
- 0.004*”number”
- 0.004*”sample”
- 0.004*”distribution”

Popular in every taxon **except** non-insect arthropods, fish, and fungi.

Some Cool Topics

Topic 64: molecular phylogenetics

- 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”

Among the **top-20 most significant probabilities** in reptiles and amphibia, birds, fish, fungi, and mammals; top-5% in every other group

How about disagreement?

Close reading of a number of papers where we know that taxonomic disagreement is taking place

How about disagreement?

Eaxmple: the “disagreement” list:

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

How about disagreement?

In the end, we prepared four lists: terms referring to **epistemic values**, **disagreement**, **pejorative evaluation**, and more general **taxonomic change**

How about disagreement?

Ask the topic model: what topics are likely to select words from our lists of disagreement and related terms?

How about disagreement?

Ask the topic model: what topics are likely to select words from our lists of disagreement and related terms?

- **Disagreement:** Topic 43
- **Epistemic values:** Topic 91
- **Pejorative terms:** Topics 43 and 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”

The terms you use to **present a new species** and to **discuss whether a species is a synonym**

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”

The terms you use to **argue about ranking of a clade**

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? 🤔

More precision?

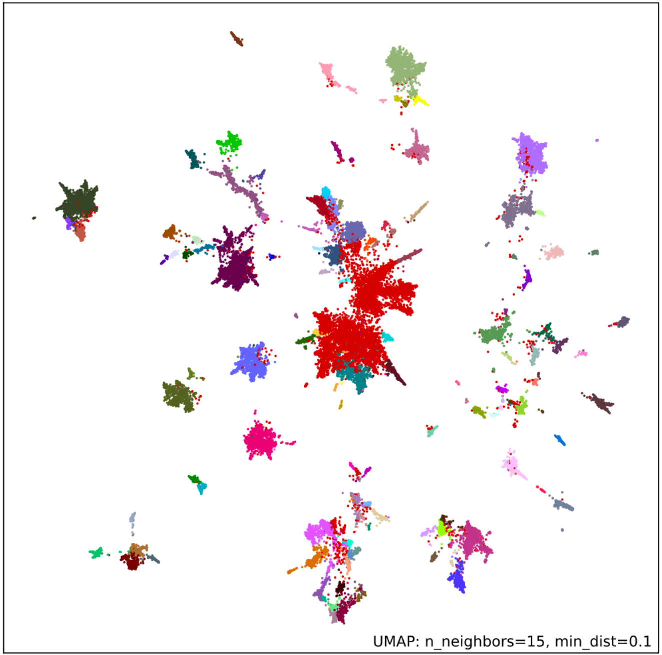
It’d be nice to distinguish between more precise uses of the kinds of terms in these topics—e.g., between **describing new species** and **declaring species to be synonyms**

Document Vector Model

Train a model that represents the words in our corpus using vectors in a 100-dimensional space,[1] and then represent each document as a vector within that same space.[2]

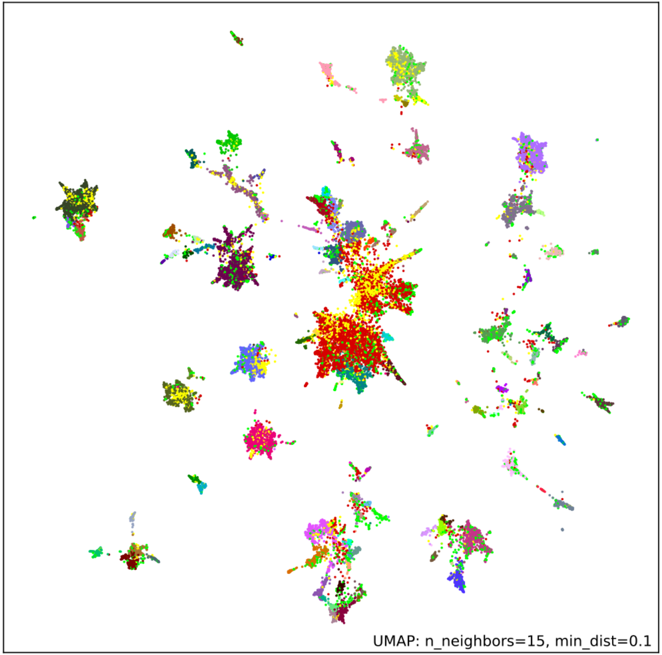
[1] technically: a Word2Vec model using hierarchical softmax

[2] technically: a Doc2Vec model, which infers vector representations of documents by sampling a sliding window of words


UMAP: n_neighbors=15, min_dist=0.1

Finding disagreement

Then: represent our disagreement terms as vectors within this space, and find the documents that are located "closest" to them!


UMAP: n_neighbors=15, min_dist=0.1

Disagreeing about what?

Which taxa are you more likely to discuss in papers that are in the "disagreement" area of the vector space? Extract all species names[3] from the top 5,000 and bottom 5,000 documents, and compare relative risk.

[3] technically: using the excellent `gnfinder` package

Disagreement by taxon

More disagreement:

Mammals (≈ 4), Birds (3), Fungi (3), Fish (2)

Less disagreement:

Insects (≈ 0.5)

Talking about disagreement

Other than disagreement words, what words distinguish the "disagreement" papers from the "non-disagreement" papers?[4]

[4] technically: apply the Craig Zeta algorithm to the top-5,000 and bottom-5,000 documents

Talking about disagreement

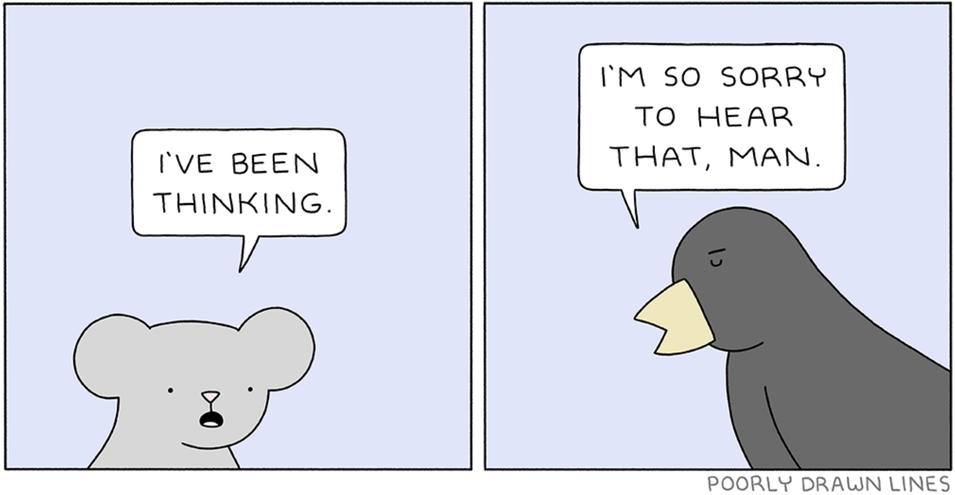
Disagreement:

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

Non-Disagreement:

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

Analysis versus Cartography


I'VE BEEN THINKING.
I'M SO SORRY TO HEAR THAT, MAN.
POORLY DRAWN LINES

Questions?

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

PENCE LAB

fnrs
LA LIBERTÉ DE CHERCHER