Screening articles for tortured phrases with a regular expressions-based detector

Alexandre Clausse, Guillaume Cabanac, Pascal Cuxac, Ophélie Fraisier-Vannier, Cyril Labbé

Screening articles for tortured phrases with a regular expressions based detector

Alexandre Clausse[1], Guillaume Cabanac[1,2], Pascal Cuxac[3], Ophélie Fraisier-Vannier[1], Cyril Labbé[4]

[1] *IRIT UMR 5505, University of Toulouse, Toulouse, France* [2] *Institut Universitaire de France, Paris, France* [3] *INIST-CNRS UAR 67, Vandoeuvre-lès-Nancy, France* [4] *LIG UMR 5217, Univ. Grenoble Alpes, Grenoble, France*

Objective	Provide a free-to-use tortured phrases detector to identify problematic articles
Contribution	A standalone algorithm using regular expressions on a list of identified fingeprints
Results	Comparable performance to the current PPS-Dimensions screening process

7k+ fingerprints | **100k**+ Hindawi articles

Experiment

Between **2020** and **2022** | **18** disciplines (STEM)

Detection of tortured phrases in the PPS

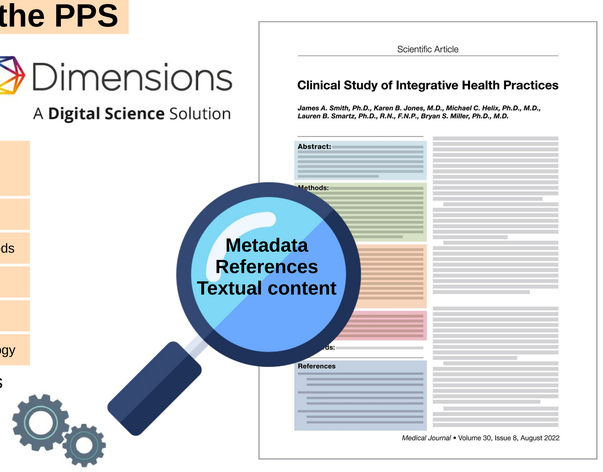
Problematic Paper Screener
Est. February 27th, 2021

Dimensions
A **Digital Science** Solution

- Detectors:
- 0 Annulled
- 1 **Tortured**
- 2 SCIgen
- 3 Mathgen
- 4 SBIR
- 5 Suspect
- 6 Seek&Blastn
- 7 Problematic Cell Lines
- 8 Citejacked
- 9 Feet of Clay

Fingerprint Tortured phrase	Expected text
“Information mining”	Data mining
“Grouping methods”	Classification methods
“Huge information”	Big data
“Crown epidemic”	Corona epidemic
“Data technology (IT)”	Information technology

7k+ fingerprints over **130M** articles



PUBPEER
The online Journal club

This Wiley article cited 7 times contains several **tortured phrases** that make some passages hard to parse. These typically result from an attempt to avoid plagiarism detection using a paraphrasing software. So far, the following have been spotted:

Tortured Phrases (found)	Established Phrases (expected)
Fourier change	Fourier transform
malignant growth AND cancer	cancer
receptive oxygen species	reactive oxygen species
sub-atomic weight	molecular weight
warmth exchange	heat transfer
statically significant	A typo: statistically significant

Can the author explain why they departed from the established phrases?

Screening process

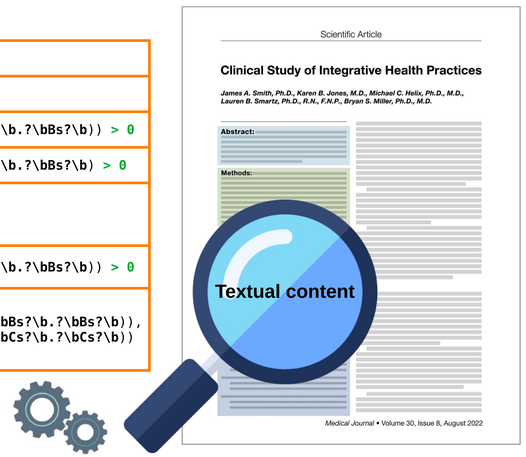
Detection of tortured phrases in our approach

Fingerprint Tortured phrase	Expected text	Fingerprint type
“Man-made consciousness”	Artificial intelligence	A
“Profound learning” AND “deep learning”	Deep learning	A AND B
“128 pieces”~5	128 bits	A B~X
“Concentrate highlight” NOT specular	Extract feature	A NOT B
“Affirmed recuperated demise” OR “affirmed recuperated passing”	Confirmed, recovered and death (cases)	A OR B
“Partial administrator” AND (“Baleanu” OR “Riemann”)	Fractional operator	A AND (B OR C)

7k+ fingerprints

Regular expression detection

```
n_matches(\bAs?\b.?\bAs?\b) > 0

min(n_matches(\bAs?\b.?\bAs?\b), n_matches(\bBs?\b.?\bBs?\b)) > 0

n_matches(\bAs?\b.?\bAs?\b (\b\w+\b ){0,X} \bBs?\b.?\bBs?\b) > 0

n_matches(\bAs?\b.?\bAs?\b) > 0
and
n_matches(\bBs?\b.?\bBs?\b) = 0

max(n_matches(\bAs?\b.?\bAs?\b), n_matches(\bBs?\b.?\bBs?\b)) > 0

max(
    min(n_matches(\bAs?\b.?\bAs?\b), n_matches(\bBs?\b.?\bBs?\b)),
    min(n_matches(\bAs?\b.?\bAs?\b), n_matches(\bCs?\b.?\bCs?\b))
) > 0
```



DOI	10.1155/2020/1280632
Tortured phrases	Brilliant AND keen AND savvy AND srewd Conveyance framework Nucleic corrosive Number juggling
Tips	4

Screening process

Results on 100k+ Hindawi articles

Hindawi XML corpus: **100k**+ articles

Flagged using regular expressions: **2,455** articles

401 articles found in Dimensions but missing from the XML corpus metadata

Flagged by the PPS: **3,400** articles

1,948 mutual results from **139** journals: **58%** overlap

Top-200 results: **48** false positives & **52** false negatives due to indexing issues

Conclusion

- We obtained comparable results to the current screening process
- We invite publishers to include our method in their evaluation process
- We provided feedback for the PPS and Dimensions
- Future work will focus on screening figures, tables, and references

References

1. Cabanac G, Labbé C, Magazinov A. The ‘Problematic Paper Screener’ automatically selects suspect publications for post-publication (re)assessment. Preprint. 2022. doi:10.48550/arXiv.2210.04895
2. Cabanac G, Labbé C, Magazinov A. Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals. Preprint. 2021. doi:10.48550/arXiv.2107.06751
3. Herzog C, Hook D, Konkiel S. Bringing down barriers between scientometricians and data. Quantitative Science Studies. 2020;1:387–395. doi:10.1162/qss_a_00020

alexandre.clausse@irit.fr irit.fr/iris

The NanoBubbles project has received Synergy grant funding from the European Research Council (ERC), within the European Union’s Horizon 2020 programme, grant agreement no. 951393.