Reproducibility review of: Semantic complexity of geographic questions - A comparison in terms of conceptual transformations of answers

Philipp A. Friese, Jakub Krukar

2023-05-30

This report is part of the reproducibility review at the AGILE conference. For more information see https://reproducible-agile.github.io/. This document is published on OSF at https://osf.io/d2shf. To cite the report use

> Friese, Philipp A. and Krukar, Jakub (2023, March). Reproducibility review of: Semantic complexity of geographic questions - A comparison in terms of conceptual transformations of answers. https://doi.org/10.17605/osf.io/d2shf

Reviewed paper

Nyamsuren, E., Xu, H., Top, E. J., Scheider, S., and Steenbergen, N.: Semantic complexity of geographic questions - A comparison in terms of conceptual transformations of answers, AGILE GIScience Ser., 4, 10, https://doi.org/10.5194/agile-giss-4-10-2023, 2023

Summary

The data and software of the paper under reproduction is published on GitHub under an MIT license. All figures, tables, and embedded data points have been reproduced. The authors showed dedication and concern to support reproducibility of their work.

Reproduction was *successful*.

Reproducibility reviewer notes

Out of the eight figures, five (4 to 8) are eligible for reproduction. Out of the three tables, one (3) is eligible for reproduction. In addition, several data points embedded into Section 5.1, 5.2.1, and 5.2.2 are eligible for reproduction. All parts have been successfully reproduced.

During reproduction, several problems arose regarding the developed Python script. These problems were mainly due to operating system differences, primarily boiling down to differing path separation symbols and differences in the execution stack underlying the used NLP packages. These issues have been resolved by changes implemented by the authors and switching to a different execution environment.

Once the issues have been resolved, reproduction was straight-forward. The reproduced tables, embedded data points, and figures are displayed at the end of this report.

Slight differences are visible in the generated data. This is expected (due to updates to the underlying third-party NLP libraries) and the differences are statistically irrelevant, according to the authors:

> *Regarding Table 3, it is expected that there is a variance in Z statistics values since the underlying data produced by NLP are slightly different as well. However, as shown by the adjusted p-values, the significant and non-significant effects remain similar to the ones reported in the article. Overall, there are no significant differences that would change the conclusions of the article.*

Comments to the authors

Reproduction was straight-forward once the technical issues have been resolved, which is commendable.

The encountered technical issues largely revolve around operating system differences. One concern raised to the authors was the use of hard-coded, operating-system dependent path separators. This concern has been addressed by the authors during reproduction. For future reproductions, I *recommend testing the developed software on several operating systems* to catch these issues. Windows users may use the WSL (1) for straight-forward access to a Linux-based execution environment.

The remaining source for technical issues regarding the execution environment may be addressed by using for example Docker or Podman containers, which may help providing a ready-to-use, encapsulated execution environment.

Data and Software Availability

All data and software developed for this reproduction is published under a CC-BY-4.0 license in the OSF repository accompanying this report at https://osf.io/d2shf.

References

[1] https://learn.microsoft.com/en-us/windows/wsl/

Reproduced Tables and Data

Table 1: Dunn's pairwise testing with Holm-Bonferroni correction on distributions of numbers of Concepts and Transformations. Corresponds to Table 3 in reproduced paper.

Corpus	Z-test C	P (adj.) C	Z-test T	P (adj.) T
Geo201-GeoAnQu	7.607	0.000	6.357	0.000
Geo201-GeoCLEF	-0.854	1.000	0.822	1.000
Geo201-GeoQuery	-1.415	1.000	1.612	1.000
Geo201-Giki	-2.392	0.168	-1.885	0.594
GeoAnQu-GeoCLEF	-5.422	0.000	-2.940	0.033
GeoAnQu-GeoQuery	-12.085	0.000	-6.805	0.000
GeoAnQu-Giki	-8.476	0.000	-6.963	0.000
GeoCLEF-GeoQuery	0.167	1.000	-0.025	1.000
GeoCLEF-Giki	-0.924	1.000	-2.086	0.370

Corpus	Z-test C	P (adj.) C	Z-test T	P (adj.) T
GeoQuery-Giki	-1.730	0.837	-3.361	0.008

Table 2: Reproduction of Kruskal-Wallis Test results for concepts and transformations. Corresponds to data in Section 5.1 in reproduced paper.

Type	Chi^2	P
Concepts	162.351	0
Transformations	74.116	0

The following two tables report Richness, Diversity, and Evenness of 'all' and 'goal' concepts, as reported by the authors in Sections 5.2.1 and 5.2.2 respectively. Note that these tables all three metrics for all corpora, not all of which are displayed in the reproduced paper. The representation in this reproduction originates from the analysis script provided by the authors in the DASA section.

Table 3: Reproduction of Richness (R), Diversity (H', Shannon Index), and Evenenss (J', Pielou Index) of all concepts. Corresponds to data in Section 5.2.1 in reproduced paper.

	Richness (R)	Diversity (H')	Evenness (J')
Geo201	10	1.390	0.604
GeoCLEF	6	0.991	0.553
Giki	8	0.769	0.370
GeoQuery	9	1.454	0.662
GeoAnQu	22	2.402	0.777

Table 4: Reproduction of Richness (R), Diversity (H', Shannon Index), and Evenenss (J', Pielou Index) of goal concepts. Corresponds to data in Section 5.2.2 in reproduced paper.

	Richness (R)	Diversity (H')	Evenness (J')
Geo201	6	0.859	0.479
GeoCLEF	4	1.016	0.733
Giki	3	0.272	0.248
GeoQuery	8	1.340	0.644
GeoAnQu	17	2.238	0.790

Reproduced Figures

This subsection displays all figures reproduced by running the `statsScript.R` file provided by the authors in their DASA section.

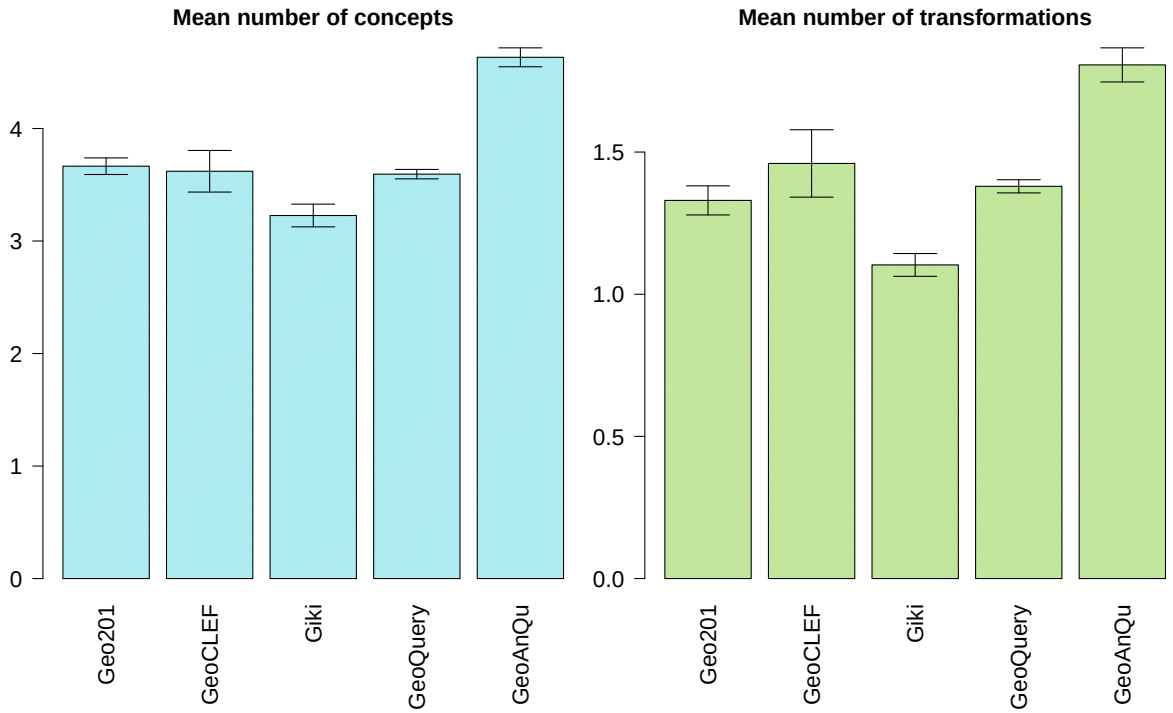

Figure 1: Mean numbers (with standard errors) of concepts and transformations per question. | Reproduction of Figure 4 in reproduced paper.

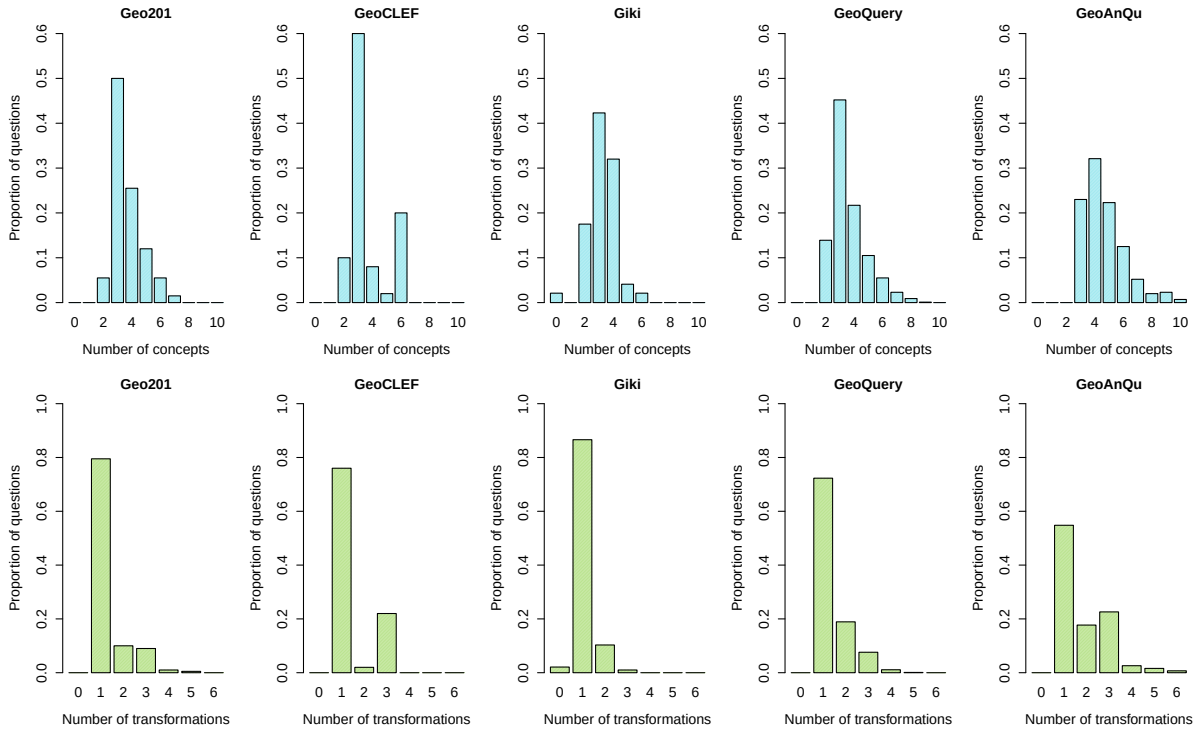

Figure 2: Transformation complexity: distributions of the number of concepts (top) and the number of transformations (bottom) per question in each corpus. | Reproduction of Figure 5 in reproduced paper,

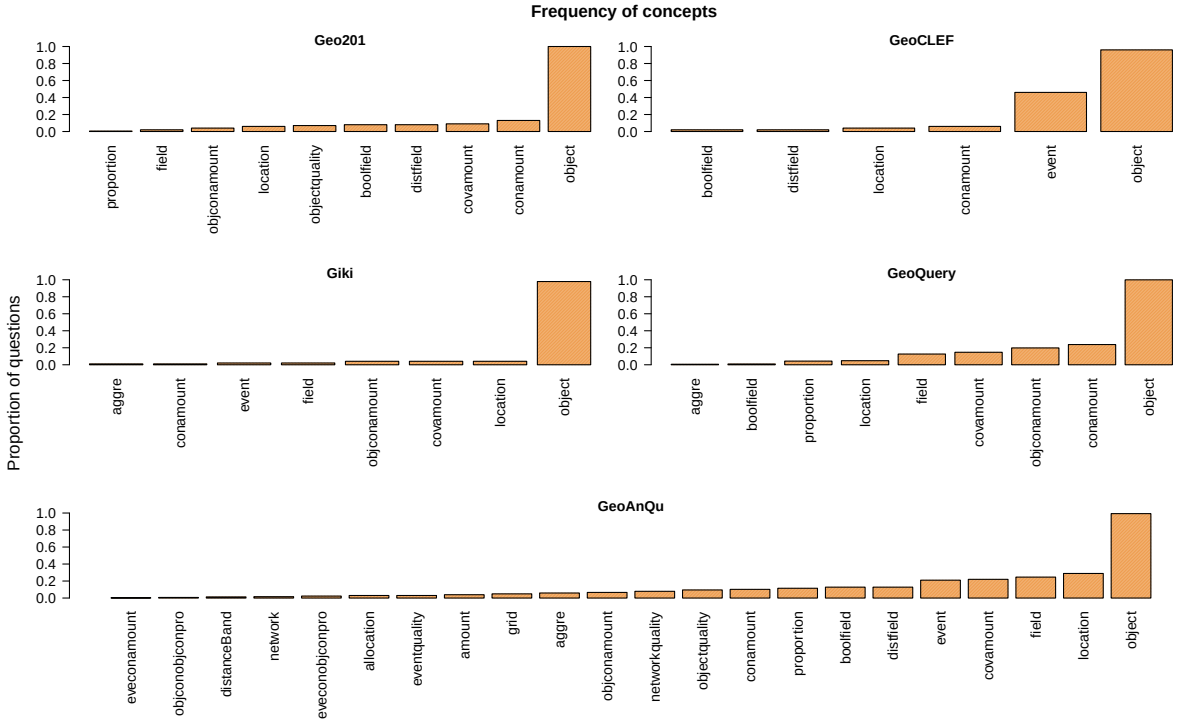

Figure 3: Frequency of all concepts as a proportion of questions where the concept occurs. | Reproduction of Figure 6 in reproduced paper.

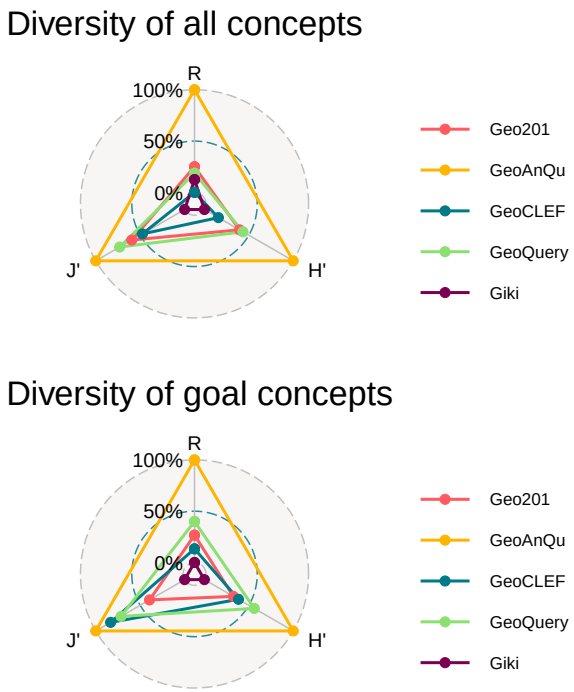

Figure 4: Diversity measures for (top) all concepts and (bottom) goal concepts. | Reproduction of Figure 7 in reproduced paper.

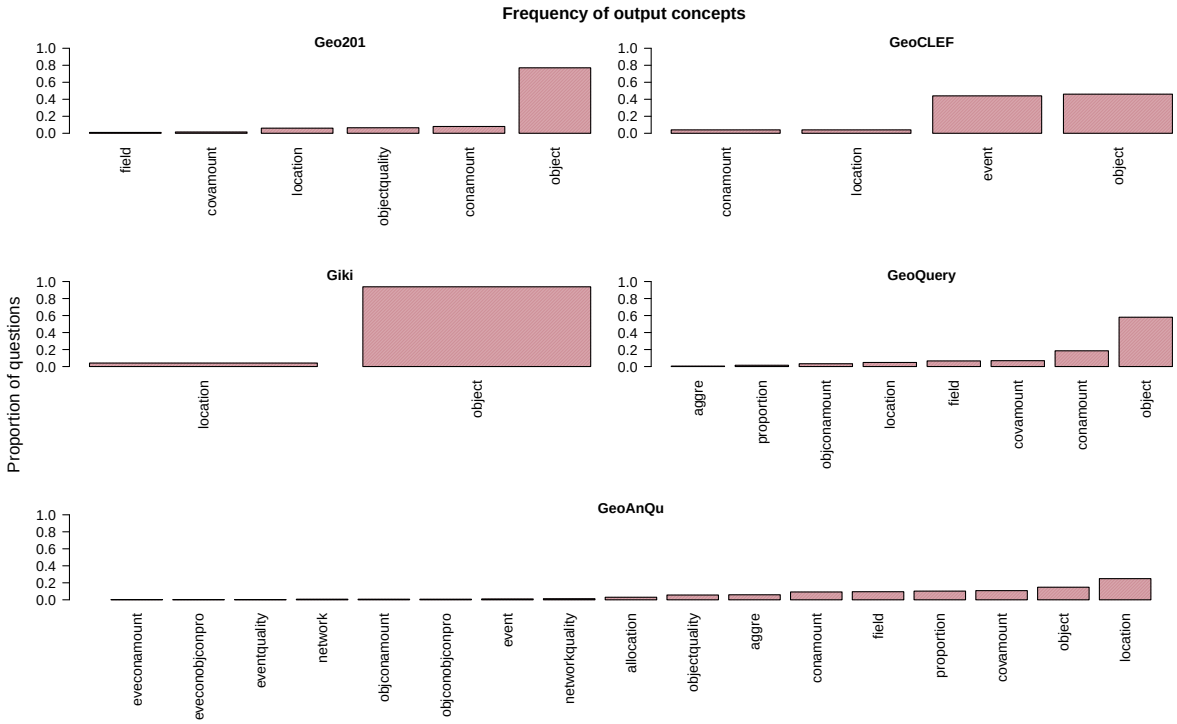

Figure 5: Frequency of goal concepts as a proportion of ques- tions where the goal concept occurs. | Reproduction of Figure 8 in reproduced paper.