AGILE: GIScience Series, 4, 10, 2023. https://doi.org/10.5194/agile-giss-4-10-2023
Proceedings of the 26th AGILE Conference on Geographic Information Science, 2023.
Editors: P. van Oosterom, H. Ploeger, A. Mansourian, S. Scheider, R. Lemmens, and B. van Loenen.
This contribution underwent peer review based on a full paper submission.


# Semantic complexity of geographic questions - A comparison in terms of conceptual transformations of answers

Enkhbold Nyamsuren[1], Haiqi Xu[1], Eric J. Top[1], Simon Scheider[1], and Niels Steenbergen[1]

[1]Department of Human Geography and Spatial Planning, Utrecht University, Vening Meineszgebouw A, Princetonlaan 8a, 3584 CB Utrecht, The Netherlands

Correspondence: Enkhbold Nyamsuren (e.nyamsuren@uu.nl)

**Abstract.** There is an increasing trend of applying AI-based automated methods to geoscience problems. An important example is a geographic question answering (geoQA) focused on answer generation via GIS workflows rather than retrieval of a factual answer. However, a representative question corpus is necessary for developing, testing, and validating such generative geoQA systems. We compare five manually constructed geographical question corpora, *GeoAnQu*, *Giki*, *GeoCLEF*, *GeoQuestions201*, and *Geoquery*, by applying a conceptual transformation parser. The parser infers geo-analytical concepts and their transformations from a geographical question, akin to an abstract GIS workflow. Transformations thus represent the complexity of geo-analytical operations necessary to answer a question. By estimating the variety of concepts and the number of transformations for each corpus, the five corpora can be compared on the level of geo-analytical complexity, which cannot be done with purely NLP-based methods. Results indicate that the questions in *GeoAnQu*, which were compiled from GIS literature, require a higher number as well as more diverse geo-analytical operations than questions from the four other corpora. Furthermore, constructing a corpus with a sufficient representation (including GIS) may require an approach targeting a uniquely qualified group of users as a source. In contrast, sampling questions from large-scale online repositories like Google, Microsoft, and Yahoo may not provide the quality necessary for testing generative geoQA systems.

**Keywords.** geographical information systems, question answering, question corpora, core concept transformation, question complexity

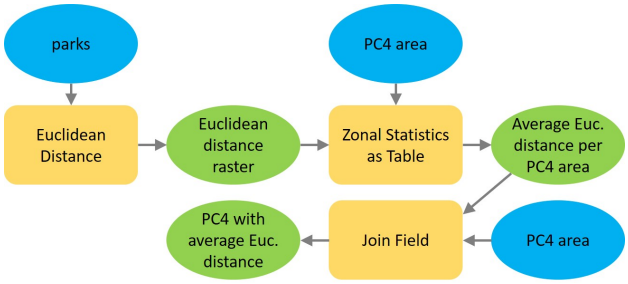
*Acknowledgements.* This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 803498).

## 1 Introduction

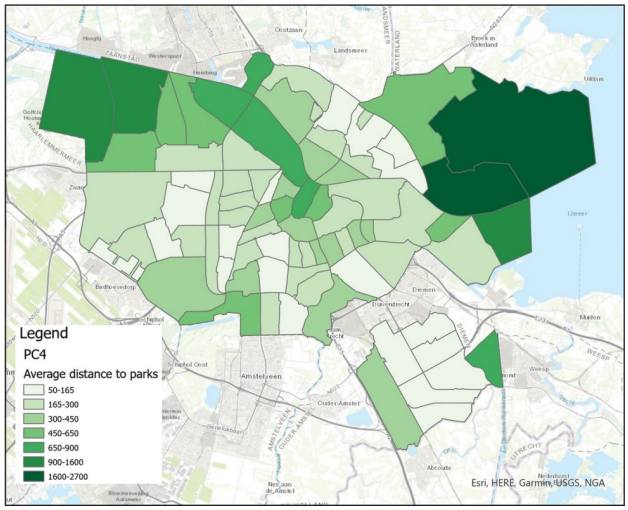
Geographic question-answering (geoQA) is gaining ground (Mai et al., 2021; Chen et al., 2021). It provides not only an Artificial Intelligence (AI) fuelled method for easing the retrieval of geographic facts from documents using natural language (such as "*What are the countries bordering Germany*"), but also a method for specifying *geographic analysis* in a way that makes sophisticated geographic information more accessible to potential users (such as in "*What is the amount of green space in Amsterdam*").

While the former task can be approached with common QA methods, the latter goes beyond the standard matching of questions to facts. Many geoQA approaches incorporate spatial query capacities for retrieval over knowledge graphs (Punjani et al., 2018). Furthermore, to handle the spatial component of geographic questions while profiting from the flexibility of machine learning models, *spatially explicit vector embeddings* (deep learning) have been recently proposed (Mai, 2021). However, geo-analytical questions are not factoid based, and thus require finding *indirect* answers, i.e., answers that need to be derived by *transforming geodata sources* with workflows (Kruiger et al., 2021). This challenge was called *geo-analytic question-answering* by Scheider et al. (2021), and interpreting questions in this way requires reasoning over *possible transformations* of *geo-analytical concepts* toward some goal.

Workflows are a common method for geo-analytical question-answering by human experts. Fig. 1a shows an example workflow constructed in ArcGIS for the 203rd question in *GeoAnQu*, "*What is the average Euclidean distance to parks for each PC4 area in Amsterdam?*". The yellow nodes represent tools. The input and output datasets are shown by blue and green nodes. Given a dataset of parks within the extent of Amsterdam, *Euclidean Distance* calculates the straight-line distance to the closest park for each cell within the extent. *Zonal Statistics As Table* aggre-




(a)




(b)

**Figure 1.** (a) Example ArcGIS workflow for question *203* in *GeoAnQu* and (b) the corresponding map visualizing the answer.

gates these distances by 4-digit postcode areas (PC4) and calculates averages. *Join Field* merges the two datasets. This final dataset is visualized as a map in Fig. 1b. The entire workflow is a series of transformations of data from one type to another, where each type represents some concept (e.g. a layer of park objects, a distance field, etc.). These transformations are discussed in more detail in section 3.1.

Compared to its syntactic complexity, the semantic complexity of a geographic question is usually very difficult to quantify. Syntactically, e.g., the two questions "*What are the countries bordering Germany*" and "*What is the amount of green space in Amsterdam*" appear to be rather similar (they are both "What" questions of similar grammatical complexity). Yet, conceptually, the first one asks only about a topological relation between countries (objects), whereas the second one includes a conversion of a landuse coverage to an amount defined using the location of a city (object). In consequence, both, the kinds of data sources and transformation steps needed to answer these questions are likely to differ in complexity. Another important distinction is that answering the latter question requires tacit expert knowledge, i.e., knowledge of the expert solving the problem rather than knowledge of the problem specification itself (McQueen, 1998; Jia et al., 2022). Such knowledge is not present in the question but comes from a human expert interpreting the question. For example, how green space could be quantified as an amount is not specified but needs to be inferred by an expert. This inference requires reinterpreting green space and amount as geo-analytical concepts and choosing applicable operations for transforming the former into the latter.

The necessity of capturing tacit knowledge for GIS workflow construction is being increasingly recognized (Liu et al., 2017; Kruiger et al., 2021). However, this raises the question of whether existing geographical question corpora are representative enough, as problem specifications, to train, test, and validate such systems (Mai et al., 2021). Are they diverse enough concerning syntactic and semantic constituents as well as the required tacit knowledge? Recently, several new corpora were proposed including *GeoQuestion201* (Punjani et al., 2018), *GeoAnQu* (Xu et al., 2020), and a subset of online geo-related queries from *MS MARCO* (Nguyen et al., 2016). Using NLP methods, Xu et al. (2020) demonstrated that the three corpora are syntactically and, to a limited extent, semantically different from each other. This was especially true for *GeoAnQu*, which contains questions typically answered with GIS workflows. Overall, the study suggested that the three corpora are distinct, and each is not representative enough by itself.

However, Xu et al. (2020)'s work can be improved by incorporating more corpora that are preferably manually constructed to ensure comparable quality. Most importantly, an NLP-based comparison method fails to capture the semantic complexity of questions concerning tacit knowledge. It does not take into account that the syntactic and terminological differences between the three corpora may be of superficial significance. For example, "*What is the number of people in Amsterdam*" and "*What is the population of Amsterdam*" are synonymous questions leading to the same answer despite syntactic differences. In this study, we aim to compare *GeoAnQu* and *GeoQuestion201* not only with a wider range of manually-constructed corpora but also by interpreting questions in a way that ignores these superficial differences.

Which concepts are needed to interpret geographic questions on this level? *Core concepts of spatial information* (Kuhn and Ballatore, 2015), including field, object, event, and network, were proposed as principal ways of conceptualizing the geographic environment. In geo-analytic QA, questions specify transformations of core concepts detailing how a question's goal can be reached (Scheider et al., 2021). Based on these insights, Xu et al. (2022) proposed a grammar to parse geographic questions in terms of *core concept transformations* forming the question interface component of a geo-analytic QA system under development (QuAnGIS). The underlying grammatical model of geographic information incorporates not only sentence structures but also tacit expert knowledge in GIS. This allows measuring the *semantic complexity* of geographic questions in terms of the complexity of transforming geographic information into an answer map.

In this paper, we use the core concept transformation grammar (Xu et al., 2022) to compare the semantic complexity of various geographic question corpora. In our analysis, we will compare the semantic and syntactic complexity of questions between 5 different corpora including.

## 2 Related Work

### 2.1 Manually Constructed Geo-Question Corpora

There are multiple manually constructed corpora for testing geographical question answering over knowledge graphs. The *Geoquery* corpus (Zelle and Mooney, 1996) consists of a database of 1000 geographical facts (states, cities, rivers, and mountains) about the United States, 880 natural language questions, and corresponding queries in Prolog language. An example question is "*What are the major cities in North Carolina?*". Each question can be answered with a fact from the database via a logical inference.

The *Giki* corpus includes 97 questions from the GikiP 2008 (Santos and Cardoso, 2008) and GikiCLEF 2009 (Santos and Cabral, 2009) tasks for answering questions requiring spatial reasoning with information from Wikipedia. The *GeoCLEF* corpus contains 50 questions generated from topics of the GeoCLEF 2005 (Gey et al., 2005) and 2006 (Gey et al., 2006) tracks for cross-language geographic information retrieval of the text. *GeoCLEF* is more focused on queries about geographical areas (e.g., "*Roman cities in Germany*") and events (e.g., "*Oil accidents in Europe*"). *Giki* queries about a wider range of entity types including food and people (e.g., "*List the basic elements of the cassata*"). Both corpora were designed for testing methods based on information retrieval.

The *GeoQuestions201* (Punjani et al., 2018) corpus contains 201 questions that can be expressed as SPARQL or GeoSPARQL queries over a linked dataset compiled from DBPedia, OpenStreetMap, and Global Administrative Areas databases. The corpus mostly contains questions requiring consideration of geospatial relation between two geographical features: "*Which counties border county Lincolnshire?*". These questions are more involved than the *Geoquery* questions since answering them requires reasoning over geometry and distances but within the capability of GeoSPARQL.

*GeoAnQu* was collected to analyze the syntax and semantics of geo-analytical questions in GIS (Xu et al., 2020). Currently, *GeoAnQu* consists of 305 questions (Xu et al., 2022) on various GIS analyses as shown in Table 1. Note that multiple analysis types can be asked in one geo-analytical question. For example, the answer to the question "*How many buildings are within 3 minutes of driving time from fire stations in Oleander?*" requires network analysis, overlay analysis, and summary statistics in GIS.

### 2.2 NLP-based Comparison of Corpora

Xu et al. (2020) compared the grammatical complexity and general intents (E.g., *what?*, *where?*, *when?*) of questions in *GeoAnQu*, *MS MARCO*, and *GeoQuestion201* corpora. The comparison relied on descriptive statistics (e.g., word count), the bag-of-words methods (word clouds, part-of-speech tagging, n-gram analysis), and semantic encoding of individual words (e.g., place type, date, entity, activity). They found that *GeoAnQu*'s questions have more complex syntactic structures than the others. Their analysis remains superficial in that only general intents and linguistic structures were analysed, not the geo-information concepts that underlie the questions. Scheider et al. (2021) claim that the automatic interpretation of geo-analytical questions requires a distinctly different approach than the automatic answering of factoid questions. Firstly, a set of concepts dedicated to geo-information, such as those by Kuhn (2012), is necessary to capture how question phrases relate to analytical assumptions. Secondly, because answers need to be procedurally generated, the request needs to be related to available data. In this paper, we thus address questions left unanswered by Xu et al. (2020), namely how the corpora's questions differ in terms of geo-semantics and transformational complexity.

### 2.3 Core Concepts of Spatial Information

Kuhn and Ballatore (2015) described spatial information and operations at a high-level view using five core content concepts. *Location* is defined as a relation that relates entities spatially. This relation can be expressed qualitatively in the example "The Netherlands contains Amsterdam", or quantitatively if you have the WGS84 coordinates of Amsterdam. Therefore, location is used to answer where questions. *Fields* represent entities that have continuous or homogeneous attributes at any position, such as temperature and land cover. It can also be understood as a continuous function between position and attribute. In practice, analysts are interested in aggregating field values into certain regions or interpolating missing values from measurements. *Objects* are individual entities that have their own identities and spatial boundaries (e.g., cities and administrative regions). The properties and relations of an object may be changed over time, but the identity remains. *Network* expresses connections between objects. Networks can tell whether two objects are connected, what the travel cost is between them, or the volume and direction of some flows between them. *Event* is a temporally bounded entity distinguished by its own identity, such as hurricanes and earthquakes. Fields, objects, and networks participate in events and are often changed by events.

Although the core content concepts can interpret spatial information in geo-analytical questions, to generate concrete transformations in practice, Xu et al. (2022) proposed new concepts and sub-concepts based on the current theory. The concept *Amount* quantify core concept or their

**Table 1.** GeoAnQu corpus

| Analysis type | GeoAnQu question example |
|---|---|
| overlay analysis | *Which buildings were affected by tornadoes in Oleander?* |
| proximity analysis | *What areas are within 300 meters of runways in Schiphol airport?* |
| density analysis | *What is the point density of cycling destinations in the Metro Vancouver region in Canada?* |
| network analysis | *What is the network distance to primary schools for children between 4 and 12 in Rotterdam?* |
| geographic pattern analysis | *Where are the clusters of fire alarms with similar priority for each 300-meter distance band in Fort Worth?* |
| areal interpolation | *What is the proportion of people over 65 for each PC4 area in Amsterdam?* |
| location-allocation analysis | *Where is the best site for a new landfill in the UK?* |

qualities and it can be subdivided into two sub-concepts: 1) *content amount* counts the number of core concepts or aggregates core concept and their quality via mean, sum, median, etc. For example, the average temperature or median household income; 2) *coverage amount* measures the spatial "coverage" of core concepts, including the size of core concepts (e.g., area of parks) and spatial distributions (e.g., a cluster of traffic accidents). The concept of *Proportion* is defined as a ratio between quantity obtained from amount. Based on the combinations of amounts and the core content concepts involved, proportions can be further subdivided. Table 2 summarizes the common sub-concepts of proportion, as well as other common sub-concepts introduced in Xu et al. (2022).

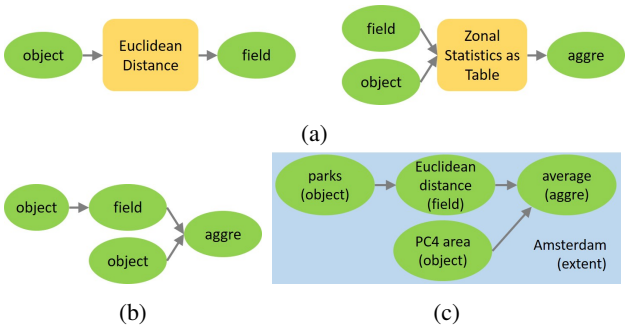



**Figure 2.** (a) Concept-based annotations of inputs and outputs for *Euclidean Distance* and *Zonal Statistics as Table* tools. (b) Example of possible transformations of concepts with *aggre* as the goal concept. (c) Transformations of the concepts from question *203*.

## 3 Parser Application

### 3.1 Concept Transformations Theory Underlying the Parser

First, each geo-analytical tool can be annotated in terms of concepts it takes as input and produces as output (Kruiger et al., 2021). For example, Fig. 2a shows annotations of the two tools from the workflow in Fig. 1a. *Euclidean Distance* transforms an *object* into a *field*. *Zonal Statistics As Table* requires *field* and *object* to output *aggre*. These two tools used in conjunction imply a series of transformations (Fig. 2b). Second, transformations of concepts can be inferred from questions with the use of expert knowledge (Xu et al., 2022). Question *203* from *GeoAnQu* is annotated as "*What is the* $average^{(aggre)}$ $Euclidean\ distance^{(field)}$ *to* $parks^{(object)}$ *for each* $PC4\ area^{(object)}$ *in* $Amsterdam^{(extent)}$". Consecutively, these concepts can also be ordered into a sequence of transformations (Fig. 2c). Here, *aggre* is the goal concept in the question and also the final node in the transformation graph. Concept-wise, transformations in Fig. 2c are the same as in Fig. 2b. This way concept transformations inferred from a question can be, at least in theory, matched to transformations of different workflows to find a workflow that can answer the question. Note that transformations in Fig. 2c do not necessarily correspond one-to-one to actual geo-analytical operations. A transformation is agnostic to the exact geo-analytical operation that may implement it and may require multiple operations.

### 3.2 Parser Implementation

Xu et al. (2022) proposed a grammar-based transformation parser for automatically annotating an English question sentence with the concepts listed in Table 2 and deriving the order of transformation of these concepts. The parser was extensively tested on the *GeoAnQu* corpus. For example, the graph in Fig. 2c was produced by this parser. Here, we briefly discuss relevant aspects of the parser and refer to the original work for detailed review (Xu et al., 2022). The parser employs the core concepts that are general and applicable to a wide range of analytical resources not limited to GIS or *GeoAnQu*. In conjunction with the constituent parts of the question sentence, these concepts encode the expert knowledge necessary to interpret and decompress a question sentence into an abstract workflow, aka concept transformations. Hence, the parsed outcome more closely reflects the analytical operations or tools than the question sentence by itself. Applying this parser to other corpora lets us compare these corpora with *GeoAnQu* on a more analytical level than would have been possible with NLP methods such as those used in Xu et al. (2020). However, the implementation of the parser relies on ANTLR-based grammar that is semantically expressive

**Table 2.** Sub-concepts proposed by Xu et al. (2022)

| Sub-concept | Explanation | Example |
|---|---|---|
| objconamount | a content amount counted from objects | population |
| eveconamount | a content amount counted from events | total votes |
| aggre | a content amount aggregated from fields, objects, networks, events or their qualities | average walking distance |
| objconobjconpro | a proportion between two objconamounts | mortality rate |
| eveconobjconpro | a proportion between an eveconamount and an objconamounts | crime rate |
| objectquality | quality of an object | age |
| networkquality | quality of a network | walking distance |
| eventquality | quality of an event | wind speed |
| distfield | a distance field generated from fields, objects, networks, events | 50-meter buffer areas of schools |
| boolfield | a Boolean field representing true or false value | within 50 meters of schools |
| grid | a sub-concept of field, proposed for normalizing geography in GIS analysis and visualization | 100 by 100-meter grid |
| allocation | a sub-concept of location, proposed for location-allocation questions | the best site |
| distanceBand | a cut-off distance for determining neighbors of each feature in cluster analysis | 300-meter distance band |

but with a limited set of syntactic patterns. This introduces restrictions on the application of the parser as discussed in the next subsection.

### 3.3 Applying the Parser to Corpora

To compare the geo-analytical complexity of the five corpora, *GeoAnQu*, *Geoquery*, *Giki*, *GeoCLEF*, and *GeoQuestions201* (further referred to as *Geo201*), we applied the transformation parser (Xu et al., 2022) on each question to (1) identify the concepts within the question and (2) generate a sequence of transformations of these concepts.

Beforehand, however, many questions in the *Giki*, *Geo201*, *GeoQuery*, and *GeoCLEF* corpora were revised for three reasons[1]. The first and main reason is malformed question sentences found in three corpora except for *Geo201*. Sentences without the interrogative pronouns (*Wh* words), such as "*Shipwrecks in the Atlantic Ocean*", are common occurrences. These questions were revised to have well-formed interrogate sentences. Other malformed questions had imperative sentences instead of interrogative sentences, such as "*Count the states which have elevations lower than what Alabama has*" in *GeoQuery*. Such sentences were revised to assume an interrogative form.

The second reason for the revision was the wordiness of the questions. These are the questions that contained unnecessary words or could have been expressed in a more laconic way. For example, "*How many people live in the biggest city in New York state*" can be rephrased into a more compact form "*What is the population of the biggest city in New York state*".

The final reason for revision is the ambiguity of the questions. For example, the goal in the question "*Rice imports in Japan*" (*GeoCLEF*) may refer to the absolute volume, the total price value, or the proportion relative to the consumption. The construction of a workflow requires a clearly stated goal in the question. Depending on this goal, the resulting workflows can be very different. Therefore, the goals in the ambiguous questions were concertized, for example, "*What is the volume of rice imports in Japan*".

After these revisions, all questions were parsed by the concept transformation parser. A quality check revealed that some questions were parsed incorrectly due to the parser's inability to recognize certain syntactic patterns and, consecutively, concepts within these structures. The parsing results for 27, 17, and 241 questions were incorrect in *Geo201*, *Giki*, and *GeoQuery* respectively. These questions were parsed manually to identify concepts and transformations.

## 4 Statistical Methods

With the parser, we have identified the concepts within the questions and the sequences of transformations of these concepts. Consecutively. the corpora were compared according to transformation complexity and diversity of transformed concepts (Fig. 3). Transformation complexity approximates the number of geo-analytical operations needed to answer a question. If questions in one corpus require more operations than questions in another corpus then we consider the former corpus more analytically complex. The diversity of transformed concepts is a proxy for the types of geo-analytical operations involved. A corpus that involves a wider range of geo-analytical operations is considered more complex and representative than a corpus that relies on repeated use of a limited set of operations.

### 4.1 Metrics for Geo-analytical Complexity

Two metrics were used for transformation complexity: the number of transformations and the number of concepts.

[1] To support reproducibility, the data distribution for this study includes both the original and revised versions of the questions. See section 4.2.

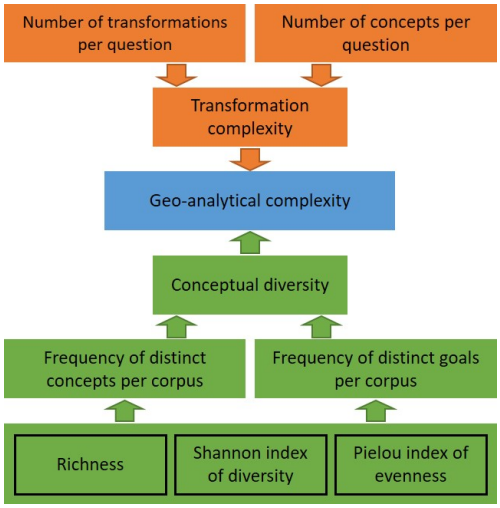



**Figure 3.** Metric used to evaluate the geo-analytical complexity of the corpora.

For example, question *203* results in two transformations and five concepts (Fig. 2c). A corpus is considered to be geo-analytically more complex if it has, on average, a higher number of transformations per question. Similarly, a more complex corpus has on average a higher number of concepts per question. As discussed, transformations do not correspond one-to-one to actual geo-analytical operations. Therefore, the number of concepts is used as the second proxy metric since a higher number of concepts correlates with more operations. Nevertheless, the number of transformations is a preferred metric as a more direct measure of geo-analytical operations (Kruiger et al., 2021).

As described previously, a transformation of concepts is agnostic to the exact geo-analytical operation(s) that may implement the transformation. Consequently, a corpus with a higher number of transformations may be limited to a few operations repeatedly used among the questions, while another corpus may involve fewer transformations per question but a higher variety of geo-analytical operations between the questions. Therefore, we used the diversity of transformed concepts as an indirect measure of the diversity of geo-analytical operations. The diversity of concepts of a corpus is represented by two metrics: (1) the frequency of distinct goal concepts in a corpus and (2) the frequency of individual concepts per corpus. Such frequency represents the number of questions in which the concept occurs. A goal concept represents the type of a question's goal, the explicit intention of a user stating the question. A corpus with a wider range of goal concepts may reflect a wider range of user intentions and operations. The second metric represents intermediary concepts necessary for achieving the goal concepts and, hence, the intermediary operations as well. Since the diversity of concepts is a function of both the number of distinct concepts and the frequency of such concepts, each metric is broken into three measures widely used for diversity estimation: richness, diversity index, and evenness. Richness is the number of distinct concepts in a corpus. Unlike richness, the Shannon-Wiener diversity index (aka Shannon index) accounts for frequencies of distinct concepts. The corpus with the higher index is the more diverse corpus. The Shannon index is estimated according to Eq. (1), where $C$ is a set of distinct concepts in a corpus, $p_i$ is a proportion of the concept $i$'s frequency relative to the sum of frequencies of all concepts in $C$.

$$H' = -\sum_{i \in C}^{C} p_i \ln(p_i) \tag{1}$$

$$J' = H'/\ln(S) \tag{2}$$

Finally, we calculate the Pielou index for evenness to obtain a more explicit measure of how much each corpus is dependent on a dominant concept. The Pielou index is calculated as in Eq. (2), where $S$ is the size of $C$. When all concepts in a corpus occur in an equal number of questions, the index is equal to one. If a corpus mostly depends on one dominant concept, the index approaches zero. Therefore, a corpus with a higher Pielou index is preferred for diversity.

### 4.2 Data and Software Availability

The concept transformation parser was implemented in Python and produces a JSON object for each question it parses. This object lists the concepts and corresponding transformations. Descriptive statistics for the JSON objects are supplied to R code for the further analysis reported in this study. All source code and datasets used/produced by this code are available on GitHub: https://github.com/quangis/AGILE2023-Semantic-complexity-GeoAnQu. The repository also provides detailed instructions for replicating the study.

## 5 Results

### 5.1 Transformation Complexity

Fig. 4 shows the mean numbers of concepts and transformations per question for each corpus. For example, a question in the *Giki* corpus has, on average, three concepts that result in one transformation, where two concepts are likely inputs, and one concept is the output (also the goal) of the transformation. The highest means of both concept and transformation counts are observed in *GeaAnQu*. This merits a more detailed analysis of the underlying distributions.

The normalized distributions as proportions of questions are shown in Fig. 5. For both metrics, the distributions of the *GeoAnQu* corpus demonstrate lower modes and longer

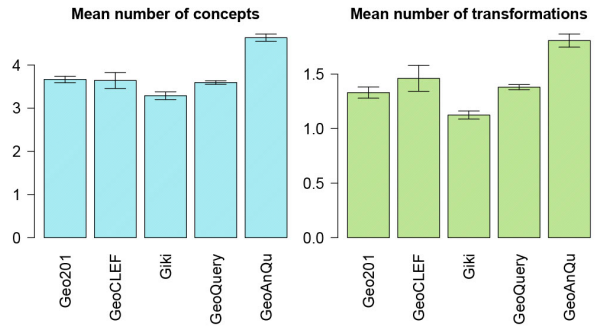



**Figure 4.** Mean numbers (with standard errors) of concepts and transformations per question.

positive skewness compared to the distributions of the other corpora. For each metric, the five distributions were analyzed for identicality with a Kruskal Wallis Test. This test is an alternative to one-way ANOVA for cases with non-normal distributions and uneven sample sizes. The tests indicate significant differences between the distributions: $H(4) = 162.35, p < .01$ for the number of concepts and $H(4) = 74.12, p < .01$ for the number of transformations.

We did a follow-up pairwise comparison of the distributions with Dunn's test with the Holm–Bonferroni correction for multiple testing. The results are summarized in Table 3. Column $p.adj$ shows p-values after the correction. In terms of the number of concepts, the *GeoAnQu* corpus has significantly more concepts than the other four corpora. No other pair of distributions have a significant adjusted p-value. A similar effect is observed for the number of transformations. *GeoAnQu* has significantly more transformations than any other corpora as strongly suggested by the p-values.

**Table 3.** Dunn's pairwise testing with the Holm–Bonferroni correction on distributions of numbers of concepts and transformations.

| | concepts | | transformations | |
|---|---|---|---|---|
| corpus1-corpus2 | z-test | p.adj | z-test | p.adj |
| Geo201-GeoAnQu | 7.61 | < 0.01 | 6.36 | < 0.01 |
| Geo201-GeoCLEF | -0.85 | 1 | 0.82 | 1 |
| Geo201-GeoQuery | -1.42 | 1 | 1.61 | 1 |
| Geo201-Giki | -2.39 | 0.17 | -1.89 | 0.59 |
| GeoAnQu-GeoCLEF | -5.42 | < 0.01 | -2.94 | 0.03 |
| GeoAnQu-GeoQuery | -12.1 | < 0.01 | -6.80 | < 0.01 |
| GeoAnQu-Giki | -8.48 | < 0.01 | -6.96 | < 0.01 |
| GeoCLEF-GeoQuery | 0.17 | 1 | -0.03 | 1 |
| GeoCLEF-Giki | -0.92 | 1 | -2.09 | 0.37 |
| GeoQuery-Giki | -1.73 | 0.84 | -3.36 | < 0.01 |

## 5.2 Conceptual Diversity

### 5.2.1 Base Diversity of Concepts

First, we review the frequency of all concepts (goal and intermediate) that were identified in questions of each corpus as shown by the distributions in Fig. 6. In each distribution, the concepts are ordered by frequency. For comparability, the frequencies are shown as the proportions of the questions in which the concept was identified. For example, the concept *event* was identified in 46% of all questions of the *GeoCLEF* corpus. However, these are not the proportions used for calculating the Shannon index of diversity. Nevertheless, these distributions provide a good context in which we can interpret the measures of richness ($R$), diversity ($H'$), and evenness ($J'$) as shown by the top chart in Fig. 7. The values in the figure are normalized with the maximum value in each scale.

Several observations can be made from the distributions of richness. First, the *GeoAnQu* corpus demonstrates the highest richness of concepts. Furthermore, every concept that occurs in one of four corpora also occurs in the *GeoAnQu* corpus. Next, not every concept that occurs in the *GeoAnQu* corpus occurs in another corpus. Following concepts occur only in the GeoAnQu corpus: *amount*, *eveconamount*, *distanceband*, *allocation*, *network*, *grid*, *objconobjconpro*, *eveconobjconpro*, *eventquality*, *networkquality*. On the other hand, only three concepts, *object*, *location*, and *conamount*, occur in all five corpora. Four other concepts, *field*, *covamount*, *boolfield*, and *objconamount* occur in three corpora apart from *GeoAnQu*.

According to the Shannon index, the *GeoAnQu* corpus has a greater diversity of concepts (2.402) than any other corpus even after taking into account the frequencies of the concepts. The *GeoQuery* corpus has the second-highest diversity index (1.454), but it is still considerably lower than for *GeoAnQu*. However, the Shannon index should be interpreted with care since each distribution in Fig. 6 exhibits the *object* as a dominant concept and a long tail of other low-frequency concepts. According to Pielou indices, *Giki* is the lowest-scoring corpus (0.37), while *GeoAnQu* still has the highest degree of evenness (0.777). Regardless, none of the corpora has an index close to one, which suggests that all corpora are biased, to a different degree, toward a few concepts with the *object* being the most dominant one in all corpora.

### 5.2.2 Diversity of Goal Concepts

Fig. 8 depicts the proportions of the questions in which the goal concept was identified. The three measures of diversity are shown in the bottom chart in Fig. 7.

Richness is lower for all corpora. That is not all concepts that are transformed occur as question goals. The *Giki* corpus had the largest decrease in richness, down to 38% from

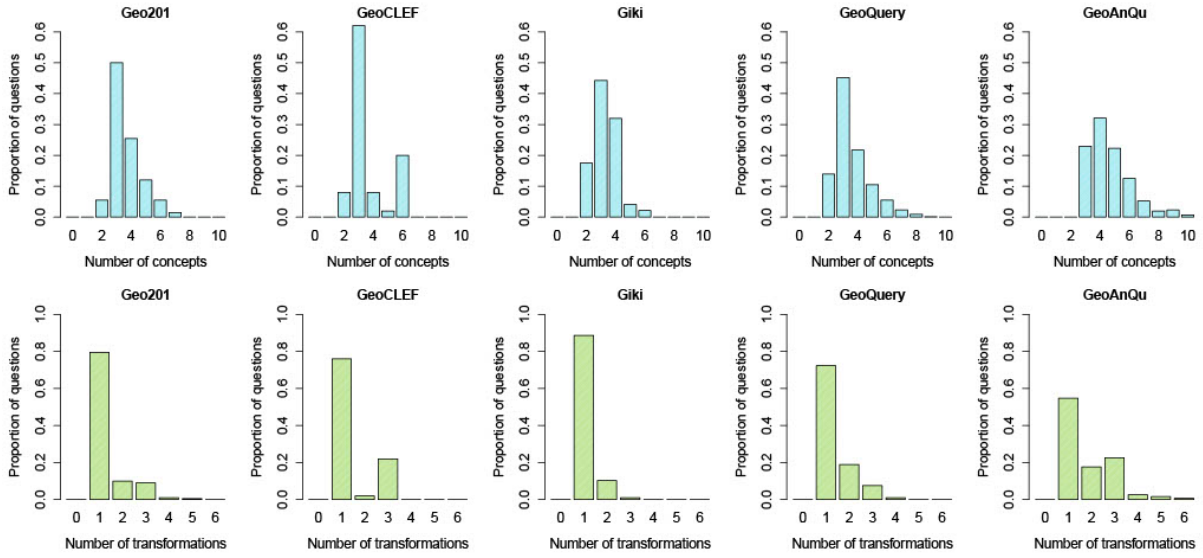



**Figure 5.** Transformation complexity: distributions of the number of concepts (top) and the number of transformations (bottom) per question in each corpus.

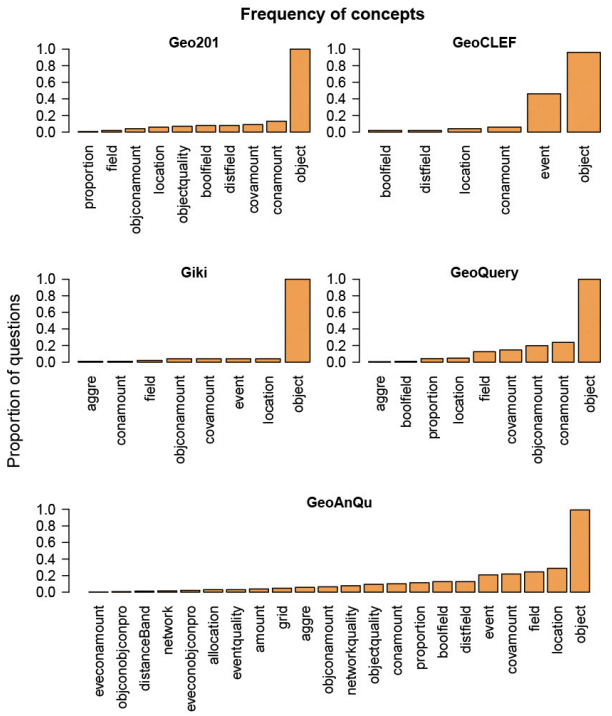



**Figure 6.** Frequency of all concepts as a proportion of questions where the concept occurs.

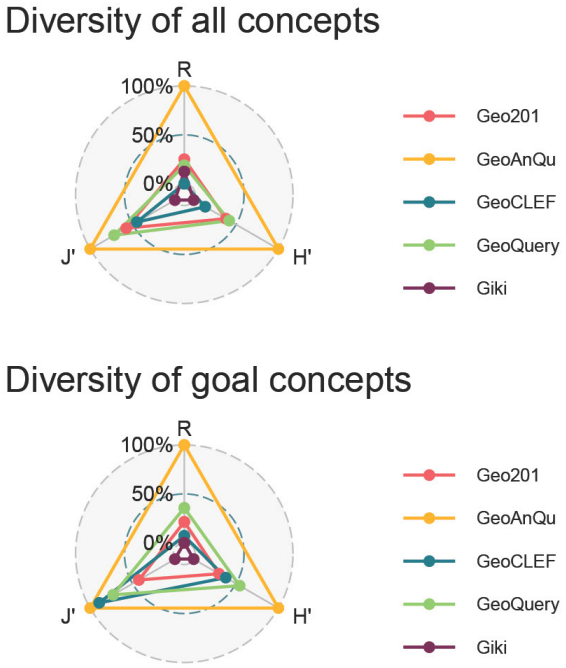



**Figure 7.** Diversity measures for (top) all concepts and (bottom) goal concepts.

all concepts, while the *GeoQuery* corpus had the least drop in conceptual richness for goals. In the *GeoAnQu* corpus, the following five concepts serve as intermediary concepts only and do not occur as a goal: *distfield*, *boolfield*, *grid*, *amount*, and *distanceBand*. Neither of these concepts occurs as a goal in the four other corpora. Despite this drop in richness, *GeoAnQu* still has the highest richness of goal concepts among the corpora.

The Shannon index is also lower for the goal concepts for four corpora. The *GeoAnQu* corpus is still the most diverse corpus concerning the goal concepts (2.242). Furthermore, it is the only other corpus with higher evenness (0.791) than its base evenness (0.777). Fig. 8 shows the goal concepts in *GeoAnQu* are more evenly distributed among the questions. The distribution has a much smoother gradient, and the *object*, as a goal concept, is much less prominent in *GeoAnQu*. In contrast, the four other corpora are

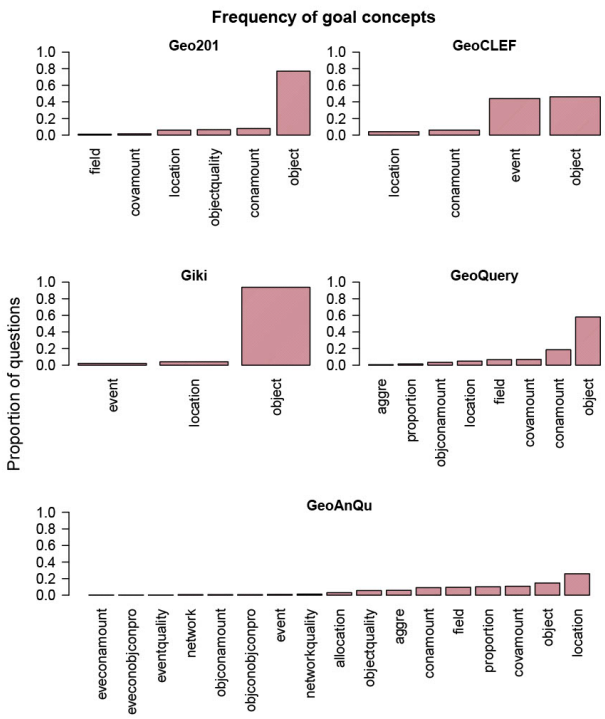



**Figure 8.** Frequency of goal concepts as a proportion of questions where the goal concept occurs.

still heavily biased toward the *object* concept. Overall, the *GeoAnQu* corpus not only demonstrates higher richness and diversity but also a more even representation of the goal concepts in the questions.

## 6 Discussion

The overall results suggest that the *GeoAnQu* corpus is considerably more complex than the other four corpora from the perspective of geo-analysis. On average, a question in the *GeoAnQu* corpus implies significantly more geo-analytical concepts than the alternatives from the four other corpora (Fig. 5 and Table 3). Similarly, the average number of transformations of these concepts is significantly higher than in any other corpus (Fig. 5 and Table 3). Both metrics suggest that answering a question from the *GeoAnQu* corpus requires more geo-analytical operations than a question from other corpora would require.

The results of conceptual diversity analysis further suggest that these geo-analytical operations are quite diverse in *GeoAnQu* and not limited to a few repeatedly used ones (Fig. 6 and Fig. 7). Since different concepts can be mapped to different operations that either take these concepts as input or produce them as output (Kruiger et al., 2021), these concepts also reflect the diversity of geo-analytical operations required by the questions in *GeoAnQu*. In terms of the richness of transformed concepts, the *GeoAnQu* corpus involves ten more concepts that are not used in any other corpus. Furthermore, the concepts from the long tail occur in *GeoAnQu* with sufficient frequency to warrant the highest values for both the Shannon index of diversity and the Pielou index of evenness. In other words, these concepts and the question that employ these concepts do not seem to be outliers or exceptions but have at least some representational value. The same can be argued for the goal concepts in *GeoAnQu* (Fig. 8 and Fig. 7). Furthermore, the goal concepts also reflect the explicit intentions of users postulating the questions. Therefore, it can also be argued that *GeoAnQu* represents a wider range of user interests than the other corpora.

By all metrics, *GeoAnQu* remains to be a more complex corpus even after comparison with a pooled corpus with questions aggregated from the four other corpora. This points toward some systematic issues within these corpora. The first issue is the syntactically different but semantically the same questions. For example, *GeoQuery* has five questions all asking about the number of cities in the US with slight variations in terminology and sentence structure. Second, some questions are syntactically the same but also don't vary much semantically. An example from *Giki* is "*Places where Goethe lived*" and "*Places where Mozart lived*". Third, the compound questions can be separated into two or more same questions. *GeoCLEF* has the following compound question "*Shark attacks near Australia and California*". An answer to such a question is the same set of operations repeatedly applied to the extents of Australia and California. Geo-analytical operations mostly revolve around identifying a limited set of topological relations such as "*A bordering B*", "*A within B*", or "*A crossing B*". This especially applies to *Geo201* and *GeoQuery*. Finally, *Giki* contains a sizeable number of questions that are rather debatable as geographic questions. The most salient examples are "*List the basic elements of the cassata*" and "*Name Portuguese-speaking Nobel prize winners*".

The results also reveal potential issues with *GeoAnQu*. The uneven distribution of intermediary concepts (Fig. 6) suggests bias in representing geo-analytical content. Although to a lesser degree than in other corpora, the questions with just one transformation are also prevalent in the *GeoAnQu* (Fig. 5). While these questions may still result in workflows with two or more operations, it needs to be verified explicitly. Otherwise, *GeoAnQu* may still be biased to "simpler" questions of those usually addressed by GIS workflows. Finally, *GeoAnQu* includes only 305 questions. Considering, the number of distinct concepts it covers, this sample size may not be enough for its purpose of building, testing, and validating geoQA systems.

## 7 Conclusion

Interpreting questions as concepts and their transformation (Xu et al., 2022) offers a novel comparison method that allowed us to take into consideration expert knowledge and compare semantics at a level less susceptible to lexical and syntactic variations. The results of this study complement

previous results from NLP-based analyses. The *GeoAnQu* corpus is distinct from other comparative corpora not only concerning syntax and semantics within the questions, as demonstrated in Xu et al. (2020), but also regarding tacit expert knowledge required for answering the questions. However, it should be noted that the higher complexity is not evidence of *GeoAnQu* being a superior alternative to the other corpora. Rather it is an indication that different corpora, including *GeoAnQu*, should be used in conjunction to comprehensively test AI systems.

Overall, the study highlights the need for a more comprehensive corpus that can represent a wider range of geo-analytical tasks that go beyond retrieval. Furthermore, it also highlights an issue in the methodology of existing corpora making them less representative of geo-analytical tasks. We postulate that it is not just a matter of adding more complex questions but finding the right knowledge source for these questions such as researchers, data analysts, and GIS experts. It is also an argument against over-reliance on big data, such as MS MARCO (Nguyen et al., 2016), to generate corpora or test geoQA systems. Due to the nature of crowd-sourcing used to generate big data (Aloteibi and Sanderson, 2014; Sanderson and Kohler, 2004), it is questionable whether such data can sufficiently represent expert groups interested in answer generation. Instead, a more targeted and knowledge-driven approach would be required to compile future corpora. While *GeoAnQu* may be a good start, more effort for a similar kind of research is needed.

## References


Aloteibi, S. and Sanderson, M.: Analyzing geographic query reformulation: An exploratory study, Journal of the Association for Information Science and Technology, 65, 13–24, 2014.

Chen, J., Tang, J., Qin, J., Liang, X., Liu, L., Xing, E. P., and Lin, L.: GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning, arXiv preprint arXiv:2105.14517, 2021.

Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., and Petras, V.: GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track overview, in: Workshop of the cross-language evaluation forum for european languages, pp. 908–919, Springer, 2005.

Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P., Di Nunzio, G. M., and Ferro, N.: GeoCLEF 2006: the CLEF 2006 cross-language geographic information retrieval track overview, in: Workshop of the Cross-Language Evaluation Forum for European Languages, pp. 852–876, Springer, 2006.

Jia, J., Zhang, Y., and Saad, M.: An approach to capturing and reusing tacit design knowledge using relational learning for knowledge graphs, Advanced Engineering Informatics, 51, 101 505, 2022.

Kruiger, J. F., Kasalica, V., Meerlo, R., Lamprecht, A.-L., Nyamsuren, E., and Scheider, S.: Loose programming of GIS workflows with geo-analytical concepts, Transactions in GIS, 25, 424–449, 2021.

Kuhn, W.: Core concepts of spatial information for transdisciplinary research, International Journal of Geographical Information Science, 26, 2267–2276, 2012.

Kuhn, W. and Ballatore, A.: Designing a Language for Spatial Computing, https://doi.org/10.1007/978-3-319-16787-9_18, 2015.

Liu, J., Liu, L., Xue, Y., Dong, J., Hu, Y., Hill, R., Guang, J., and Li, C.: Grid workflow validation using ontology-based tacit knowledge: A case study for quantitative remote sensing applications, Computers & Geosciences, 98, 46–54, 2017.

Mai, G.: Geographic Question Answering with Spatially-Explicit Machine Learning Models, Ph.D. thesis, UC Santa Barbara, 2021.

Mai, G., Janowicz, K., Zhu, R., Cai, L., and Lao, N.: Geographic question answering: challenges, uniqueness, classification, and future directions, AGILE: GIScience Series, 2, 1–21, 2021.

McQueen, R.: Four views of knowledge and knowledge management, pp. 609–611, 1998.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L.: MS MARCO: A human generated machine reading comprehension dataset, in: CoCo@ NIPs, 2016.

Punjani, D., Singh, K., Both, A., Koubarakis, M., Angelidis, I., Bereta, K., Beris, T., Bilidas, D., Ioannidis, T., Karalis, N., et al.: Template-based question answering over linked geospatial data, in: Proceedings of the 12th Workshop on Geographic Information Retrieval, pp. 1–10, 2018.

Sanderson, M. and Kohler, J.: Analyzing geographic queries, in: SIGIR Workshop on Geographic Information Retrieval, pp. 8–10, 2004.

Santos, D. and Cabral, L. M.: GikiCLEF: Expectations and lessons learned, in: Workshop of the Cross-Language Evaluation Forum for European Languages, pp. 212–222, Springer, 2009.

Santos, D. and Cardoso, N.: GikiP: Evaluating geographical answers from Wikipedia, in: Proceedings of the 5th Workshop on Geographic Information Retrieval, pp. 59–60, 2008.

Scheider, S., Nyamsuren, E., Kruiger, H., and Xu, H.: Geo-analytical question-answering with GIS, International Journal of Digital Earth, 14, 1–14, 2021.

Xu, H., Hamzei, E., Nyamsuren, E., Winter, S., Tomko, M., and Scheider, S.: Extracting interrogative intents and concepts from geo-analytic questions, in: Proceedings of the 23rd AGILE conference on Geographic Information Science. Springer, (Lecture Notes in Geoinformation and Cartography), Springer, 2020.

Xu, H., Nyamsuren, E., Scheider, S., and Top, E.: A grammar for interpreting geo-analytical questions as concept transformations, International Journal of Geographical Information Science, 0, 1–31, https://doi.org/10.1080/13658816.2022.2077947, 2022.

Zelle, J. M. and Mooney, R. J.: Learning to Parse Database Queries Using Inductive Logic Programming, in: Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2, pp. 1050–1055, AAAI Press, 1996.