Noname manuscript No.
(will be inserted by the editor)

An Evaluative Baseline for Geo-Semantic Relatedness and Similarity

Andrea Ballatore · Michela Bertolotto · David C. Wilson

Received: date / Accepted: date

Abstract In geographic information science and semantics, the computation of semantic similarity is widely recognised as key to supporting a vast number of tasks in information integration and retrieval. By contrast, the role of geo-semantic relatedness has been largely ignored. In natural language processing, semantic relatedness is often confused with the more specific semantic similarity. In this article, we discuss a notion of geo-semantic relatedness based on Lehrer's semantic fields, and we compare it with geo-semantic similarity. We then describe and validate the *Geo Relatedness and Similarity Dataset (GeReSiD)*, a new open dataset designed to evaluate computational measures of geo-semantic relatedness and similarity. This dataset is larger than existing datasets of this kind, and includes 97 geographic terms combined into 50 term pairs rated by 203 human subjects. GeReSiD is available online and can be used as an evaluation baseline to determine empirically to what degree a given computational model approximates geo-semantic relatedness and similarity.

Keywords geo-semantic relatedness · geo-semantic similarity · gold standards · geo-semantics · cognitive plausibility · GeReSiD

1 Introduction

Is *lake* related to *river*? Is *road* related to *transportation*? Are *mountain* and *hill* more related than *mountain* and *lake*? While it may seem natural to answer yes to all of these questions, the logical and computational formalisation

Andrea Ballatore & Michela Bertolotto
School of Computer Science and Informatics
University College Dublin, Belfield, Dublin 4, Ireland
E-mail: {andrea.ballatore,michela.bertolotto}@ucd.ie

David C. Wilson
Department of Software and Information Systems
University of North Carolina, Charlotte, NC, USA
E-mail: davils@uncc.edu

of why this is the case has raised considerable interest in philosophy, psychology, linguistics and, more recently, in computer science. The human ability to detect semantic relatedness is essential to perform key operations in communication, such as word-sense disambiguation (e.g. interpreting *bank* as financial institution or as the terrain alongside the bed of a river), reducing semantic ambiguity and increasing efficiency in meaning-creation and sharing. The human cognitive apparatus possesses a remarkable ability to detect co-occurrence patterns that are not due to chance, but that indicate the existence of some semantic relation between the terms.

Semantic similarity has been identified as a particular subset of this general notion of semantic relatedness. While semantically *related* terms are connected by any kind of relation, semantically *similar* terms are related by synonymy, hyponymy, and hypernymy, all of which involve an *is a* relation. In this sense, *train* and *bus* are intuitively similar (they are both means of transport), whilst *bus* and *road* are related but not similar (i.e. they often co-occur but with different roles). Semantic similarity relies on the general cognitive ability to detect similar patterns in stimuli, which attracts considerable attention in cognitive science. Notably, Goldstone and Son [15] stated that "assessments of similarity are fundamental to cognition because similarities in the world are revealing. The world is an orderly enough place that similar objects and events tend to behave similarly" (p. 13). Therefore, the vast applicability of semantic similarity in computer and information science should come as no surprise.

In geographic information science (GIScience), the theoretical and practical importance of geo-semantic similarity has been fully acknowledged, resulting in a growing body of research [2, 4, 23]. By contrast, the importance of semantic relatedness, which is widely studied in the non-geographic domain, has been almost completely ignored, with the exception of the works by Hecht and Raubal [16] and Hecht et al. [17]. Computational measures of semantic relatedness play a pivotal role in natural language processing, information retrieval (IR), and word sense disambiguation, providing access to deeper semantic connections between words and sets of words. Despite the large number of existing measures, their rigorous evaluation still constitutes an important research challenge [12].

This article contributes to GIScience and semantics in the following ways. First, we discuss in detail the notion of geo-semantic relatedness, drawing on Lehrer's theory of semantic fields, which consist of sets of terms covering a restricted semantic domain. Geo-semantic relatedness is defined with respect to specifiable geographic relations between terms, and is compared and contrasted with the more widely studied geo-semantic similarity. Second, we have developed and validated the Geo Relatedness and Similarity Dataset (GeReSiD), tackling the complex issue of the evaluation of computational measures of geo-semantic relatedness and similarity. In this new dataset, we have collected psychological judgements about 50 pairs of terms, covering 97 unique geographic terms, from 203 human subjects. The human judgements in GeReSiD focus explicitly on geo-semantic relatedness and similarity between geographic terms.

The resulting dataset provides an evaluation test bed for geo-semantic relatedness and similarity. This is compared against the existing human-generated gold standards used to assess computational measures of semantic relatedness and similarity, highlighting the limitations of such datasets. Such an evaluative baseline constitutes a valuable ground truth against which computational measures can be assessed, providing empirical evidence about the cognitive plausibility of the measures. GeReSiD can inform research in geo-semantics, indicating to what degree computational approaches match human judgements. More specifically the contribution of this evaluative baseline consists of the following aspects:

- GeReSiD covers a sample of geographic terms larger than existing similarity datasets, including 97 natural and man-made unique terms, grouped in 50 unique pairs. Psychological judgements of geo-semantic relatedness and similarity were collected separately on the 50 pairs.
- GeReSiD includes a sample of evenly distributed relatedness/similarity judgements, ranging from near-synonymity to no relationship between the terms. Our methodology is described explicitly and precisely, in order to provide practical guidelines to construct similar datasets.
- Unlike existing datasets, the semantic judgements on the term pairs contained in GeReSiD are analysed with respect to interrater agreement (IRA) and interrater reliability (IRR).
- The psychological judgements in GeReSiD can be observed as the mean of relatedness/similarity of the pairs, using correlation coefficients of relatedness/similarity rankings (such as Spearman's ρ or Kendall's τ). Alternatively, the data can be interpreted as categorical, using Cohen's kappa or Fisher's exact test [7] to evaluate the computational measure.
- GeReSiD is an open dataset freely available online.[1] Both raw data and the resulting dataset are available.

The remainder of this article is organised as follows. Section 2 discusses in depth the two key notions of geo-semantic relatedness and similarity, proposing a synthetic definition. Section 3 summarises existing datasets for semantic relatedness and similarity, with particular attention to those restricted to the geographic domain. The new evaluative baseline, GeReSiD, is outlined, analysed and discussed in Section 4. Conclusions and directions for future research are indicated in Section 5.

2 Geo-semantic relatedness and similarity

This section introduces the notion of *geo-semantic relatedness*, comparing it and contrasting it with *geo-semantic similarity*. In the natural language processing literature, several terms are used inconsistently, including semantic relatedness, relational similarity, taxonomical similarity, semantic association,

[1] `http://github.com/ucd-spatial/Datasets` (acc. Apr 10, 2013)

analogy, and attributional similarity [53]. These terms are often used interchangeably [9]. A striking example of this tendency is the article title 'WordNet::Similarity: Measuring the relatedness of terms' [39].

In natural language, terms are connected by an open set of semantic relations. Common semantic relations are synonymy (A coincides with B), antonymy (A is the opposite of B), hyponymy (A is a B), hypernymy (B is a A), holonymy (A is whole of B), meronymy (A is part of B), causality (A causes B), temporal contiguity (A occurs at the same time as B), and function (A is used to perform B). Khoo and Na [28] have surveyed these semantic relations, whilst Morris and Hirst [36] have explored other non-classical semantic relations. As Khoo and Na [28] remarked, semantic relations are characterised by productivity (new relations can be easily created), uncountability (semantic relations are an open class and cannot be counted), and predictability (they follow general, recurring patterns). In the geographic domain, spatial relations such as proximity (A is near B), and containment (A is within B) have an impact on semantics [49].

Before providing our definition of geo-semantic relatedness and similarity, it is beneficial to review the semantics of these terms in the literature. In the context of semantic networks, Rada et al. [40] suggested that semantic relatedness is "based on an aggregate of the interconnections between the terms" (p. 18). To obtain semantic similarity, the observation must be restricted to taxonomic *is_a* relationships between terms. Resnik [41] followed this approach, and defined semantic similarity and relatedness as follows: "Semantic similarity represents a special case of semantic relatedness: for example, cars and gasoline would seem to be more closely related than, say, cars and bicycles, but the latter pair are certainly more similar" (p. 448).

More recently, Turney [53] added a further distinction between 'attributional' and 'relational similarity.' Following the approach outlined by Medin et al. [32], 'attributes' are statements about a term that take only one parameter, e.g. *X is red*, *X is long*. Therefore, attributional similarity measures the correspondence between the attributes of the two terms. 'Relations,' on the other hand, are statements that take two or more parameters, e.g. *X is a Y*, *X is longer than Y*. Hence, relational similarity is based on the common relations between two pairs of terms [53]. On these assumptions, synonymy is seen as a high degree of attributional similarity between two terms, e.g. <river,stream>. Analogy, by contrast, is characterised as a high degree of relational similarity between two pairs of terms, e.g. <boat,river> and <car,road>. The next sections discuss geo-semantic relatedness and similarity in detail.

2.1 Geo-semantic relatedness

A general notion of *relatedness* in the geographic context was stated in Tobler's first law, which asserts that everything is related to everything else, but near things are more related than distant things [51]. While this law was formulated

to express intuitively the high spatial autocorrelation of many geographic phenomena, it has generated several responses in GIScience. For example, in the context of information visualisation, Montello et al. [35] have proposed the *first law of cognitive geography*, which states that "people believe closer things to be more similar than distant things" (p. 317). Applying the same intuition to the domain of geo-semantics, we assert that two terms are *geo-semantically related* to the degree to which they refer to entities or phenomena connected via specifiable relations grounded in the geographic dimension.

To define a notion of geo-semantic relatedness, we rely on the notion of *semantic field*. According to Lehrer [31], a semantic field is "a set of lexemes which cover a certain conceptual domain and which bear certain specifiable relation to one another" (p. 283). While a 'domain' is an epistemological notion referring to a subset of human knowledge and experience (e.g. geography, politics, medicine, etc.), a semantic field is a more specific linguistic notion that refers to a set of lexemes utilised to describe a domain. For example, a semantic field might be formed by terms *train*, *bus*, *trip*, *fare*, *delay*, *accident*, etc., which are all connected to the underlying term of transportation, and commonly used to generate observations on the domain of mobility.

Terms appear to be semantically related to the degree to which they belong to the same semantic field, and can indeed belong to different semantic fields. Semantic fields are neither static nor well-defined sets, but rather fuzzy configurations that shift over time, and across different agents and information communities. The condition of *specifiability* of relations emphasises the fact that random co-occurrence has no impact on semantic relatedness. If a relation is not specifiable, the co-occurrence of the two terms must be random. A term has a certain degree of *centrality* in a semantic field, i.e. the density of connectedness with other terms. For example, in the aforementioned semantic field on transportation, *car* is more central than *delay*. Similarly, in a semantic field on social life, *car* is likely to be less central than *restaurant* or *pub*.

Geo-semantic relatedness can therefore be defined as a specific sub-domain of semantic relatedness, focusing on relations grounded in the geographic dimension, i.e. relations in which at least one of the terms has a spatial dimension. Examples of geo-semantically related terms are *judge*, *trial*, and *tribunal*, where *tribunal* has a strong geographic component that grounds the other terms geographically. A computational measure of geo-semantic relatedness has to aggregate and quantify the intensity of such relations between two terms, providing a useful tool for several complex tasks. For example, terms *river* and *flood* should be more geo-semantically related than *vehicle* and *car*, which possess a less prominent geographic component. Acknowledging the fact that most terms in natural language have some degree of geographic ground, we express this approach to geo-semantic relatedness following Tobler's first law of geography:

> *Every term is geo-semantically related to all other terms, but terms that co-occur with specifiable geographic relations are more related than other terms.*

In other words, every term can in principle have some degree of geo-semantic relatedness to any other term, but terms that co-occur in observations bearing specifiable relations tend to be more geo-semantically related than those that do not. This formulation puts terms in relation to human spatial experience from which terms arise, suggesting indistinct, gradual, and shifting boundaries between geo-related and unrelated terms.

In this sense, geo-semantic relatedness is intrinsically fuzzy, admitting a continuous spectrum of relatedness rather than a binary classification (i.e. *related* or *unrelated*). Highly related terms belong to the same semantic field. The same terms can belong to several overlapping semantic fields. Relatedness involves all semantic relations, including synonymy, antonymy, hyponymy, hypernymy, holonymy, meronymy, causality, temporal contiguity, function, proximity, and containment. This law applies both to natural language, where geographic terms can be highly imprecise and vague, and to scientific conceptualisations, which generally aim at stricter semantics.

Surprisingly, in GIScience semantic relatedness has been almost completely ignored, with two notable exceptions [16, 17]. In order to explore semantically and spatially related entities in Wikipedia, Hecht and Raubal [16] developed ExploSR, a graph-based relatedness measure. ExploSR computes a semantic relatedness score of two articles by assigning weights to spatially-referenced articles in the Wikipedia Article Graph. More recently, the *Atlasify* system generates human-readable explanations of the relationship between terms to support exploratory search [17].

Geo-semantic relatedness can be informed by ideas developed in the area of text mining. The latent Dirichlet allocation (LDA) adopts a probabilistic approach to cluster highly semantically-related terms in a text corpus [8]. LDA was extended to include a geographic dimension into the Location Aware Topic Model (LATM) [55]. LATM quantifies the geo-semantic relatedness between keywords, topics, and geographic locations, adopting a fully distributional approach.

2.2 Geo-semantic similarity

While geo-semantic relatedness of terms can be based on co-occurrence in observations, geo-semantic similarity of terms can only be determined through the analysis of the terms' attributes and relations. Geo-semantic similarity is a subset of geo-semantic relatedness: all similar terms are also related, but related terms are not necessarily similar. The relations considered for geo-semantic similarity include only synonymy, hyponymy, and hypernymy. Unlike geo-semantic relatedness, geo-semantic similarity has been deeply explored by the GIScience community, and is recognised as one of the key concepts of geo-semantics [29].

Several theories of similarity have been used to conceptualise and measure geo-semantic similarity, including featural, transformational, geometric, and alignment models [47, 22, 23, 48]. Specific techniques have been devised for

specific knowledge-representation formalisms [20, 44]. More recently, graph-based [5] and lexical techniques [3, 4] have been investigated in the emerging area of volunteered geographic information (VGI). These works tend to focus on the conceptual level, computing the similarity of abstract geographic terms (e.g. *city* and *river*), rather than the instance level (e.g. *New York* and *Danube*).

Beyond the specificities of such approaches, we can state that terms A and B are semantically similar with respect to C, where C is a set of attributes and relations, also known as *context* [26]. The context C focuses on the typical spatial organisation and appearance of the entity identified by the term (e.g. shape, size, material composition). Alternatively, the similarity of A and B can be measured with respect to their *affordances*, i.e. the possibilities that an entity offers to humans [19].

As observed in relation to geo-semantic relatedness, all terms can be geo-semantically similar to some limited extent, and geo-semantic similarity is therefore best modelled as a continuous spectrum, rather than a binary classification. For example, terms *restaurant* and *continent* are similar with respect to the fact that they both refer to geographically-grounded entities. To capture this idea at the linguistic level that is relevant to this discussion, we adopt the approach outlined in [4]. Considering the terms used in lexical definitions of terms, we state recursively that:

> *All terms are geo-semantically similar, but geographic terms described using the same terms are more similar than other terms.*

A geo-semantic similarity measure has to quantify the similarity of two terms into a score, enabling a number of semantic tasks in IR and information integration. For example, terms *restaurant* and *pub* are very similar because they share similar spatial organisation and affordances. *Houses* and *schools* are geo-semantically similar with respect to their spatial organisation of parts and can be described as having walls, windows, doors, a roof, etc. *Roads* and *rivers* show similar affordances – they can be used for transportation.

3 Semantic relatedness and similarity gold standards

Semantic similarity and relatedness measures can be evaluated against a human-generated set of psychological judgements. This section gives an overview of published similarity and relatedness gold standards, mostly from psychology and computational linguistics. The term 'gold standard' is described by the Oxford Dictionary of English as "a thing of superior quality which serves as a point of reference against which other things of its type may be compared."[2] In computer science, the term is used to describe high-quality, human-generated datasets, capturing human behaviour in relation to a well-defined task. Such datasets can then be used to assess the performance of automatic approaches,

[2] `http://oxforddictionaries.com/definition/gold+standard` (acc. Apr 10, 2013)

by quantifying the correlation between the machine and the human-generated data.

3.1 Cognitive plausibility

In a seminal discussion on expert systems, Strube [50] argued that knowledge engineering should strive towards increasing the *cognitive adequacy* of computational systems, defined as their 'degree of nearness to human cognition' (p. 165). In the context of GIScience, geo-relatedness or geo-similarity measures need not replicate the workings of human mind in their entirety (defined as *absolutely strong adequacy*), but should aim at what Strube called *relatively strong adequacy*, i.e. the ability of the system to function like a human expert in a circumscribed domain. Following this approach, we adopt the notion of *cognitive plausibility* to assess to what degree a measure mimics human behaviour [27].

In order to quantify the cognitive plausibility of a computational semantic relatedness or similarity measure, two complementary approaches can be adopted: (1) psychological evaluations, and (2) task-based evaluations. In psychological evaluations, human subjects are asked to rank or rate term pairs. These rankings or ratings are then compared with computer-generated rankings, usually using correlation as an indicator of performance. Alternatively, human subjects can perform a task based on the assessment of relatedness or similarity, such as word sense disambiguation, and the cognitive plausibility of the measure is observed indirectly in the results of the task, using for example precision and recall measures. Such human-generated datasets are used as gold standards.

The usage of gold standards is common in natural language processing tasks, such as part-of-speech tagging, entity resolution, and word sense disambiguation [46, 52, 10, 38]. Adopting this approach, a technique or a model can be deemed to be more or less plausible by observing its correlation with human-generated results. Such datasets are created by combining the results from a number of human subjects who perform a given task, either under controlled conditions, or through online forms. To be considered valid by a research community, a gold standard needs to meet certain criteria, such as coverage, quality, precision, and inter-subject agreement. Disagreements about the validity of a gold standard are quite common and, when weaknesses are uncovered, a gold standard can be demoted to a golden calf [e.g. 24].

The intrinsic high subjectivity of relatedness and similarity rankings makes the collection and validation of gold standards complex and challenging. Although task-based evaluations might appear more 'objective,' they are equally affected by subjectivity: ultimately, relatedness-based or similarity-based tasks are generated, interpreted, and validated by human subjects. Acknowledging the unlikelihood of total agreement, the reliability of a similarity evaluation should be grounded in stability over time, consistency across different datasets, and reproducibility of psychological results. Ideally, both evaluation

approaches should show convergent, cross-validating results: a strong correlation is expected between the cognitive plausibility of a measure and its performance in similarity-based tasks.

3.2 Comparison of relatedness and similarity gold standards

Over the past 50 years, several authors investigating semantic issues in psychology, linguistics, and computer science created datasets focused on semantic similarity and, more recently, semantic relatedness. The first similarity gold standard was published in 1965, in a article in which Rubenstein and Goodenough [45] collected a set of 65 word pairs ranked by their synonymy. Following a similar line of research, Miller and Charles [33] published a similar dataset with 30 word pairs in 1991. More recently, Finkelstein et al. [13] created the WordSimilarity-353 dataset, which contains 353 word pairs actually ranked by semantic relatedness.[3] The dataset was subsequently extended to distinguish between similarity and relatedness [1].[4] In a study of the retrieval mechanism of memories, Nelson et al. [37] collected associative similarity ratings for 1,016 word pairs.

A smaller number of geo-semantic similarity datasets have been generated in the areas of GIScience and geographic information retrieval (GIR). In this area, Janowicz et al. [21] conducted a study on the cognitive plausibility of their Sim-DL similarity measure. However, the study was conducted in German on a very small set of terms, and for this reason it is difficult to reuse in different contexts. In order to evaluate their Matching-Distance Similarity Measure (MDSM), Rodríguez and Egenhofer [44] collected similarity judgements about geographic terms, including large natural entities (e.g. *mountain* and *forest*), and man-made features (e.g. *bridge* and *house*). Before GeReSiD, the MDSM evaluation dataset was the largest similarity gold standard for geographic terms. For this reason, this dataset was utilised to carry out the evaluation of network-based similarity measures [5]. In contrast, geo-semantic relatedness has been largely ignored in the geospatial domain.

The salient characteristics of these gold standards are summarised in Table 1, detailing their human subjects, the terms and term pairs. For each dataset, the table shows whether they focus on semantic relatedness (REL), semantic similarity (SIM), and exclusively on the geographic domain (GEO). The existing datasets are compared with GeReSiD, the gold standard described in Section 4, and have several limitations. First, the procedure followed to construct the datasets is usually only sketched and not described in detail. Second, the size of the datasets tends to be rather small.

The size of such datasets can be observed along three dimensions: number of human subjects, number of terms, and number of term pairs. A clear trade-off exists between number of human subjects and number of term pairs.

[3] `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353` (acc. Apr 10, 2013)

[4] `http://alfonseca.org/eng/research/wordsim353.html` (acc. Apr 10, 2013)

Reference	Subjects	Terms & term pairs	REL	SIM	GEO
Rubenstein and Goodenough [45]	51 paid college undergrads: group I (15 subjects), group II (36 subjects).	48 terms (ordinary English words); 65 term pairs, ranging from highly synonymous pairs to semantically unrelated pairs.		√	
Miller and Charles [33]	38 undergraduate students. US English native speakers.	40 terms selected from Rubenstein and Goodenough [45]; 32 term pairs.		√	
Finkelstein et al. [13], Agirre et al. [1]	13 experts for first set (153 pairs), 16 experts for second set (200 pairs). Near-native English proficiency.	346 terms (manually selected nouns and compound nouns); 353 term pairs.	√	√	
Rodríguez and Egenhofer [44]	72 paid undergrad students (two groups of 36 people). US English native speakers.	33 geographic terms from WordNet and SDTS; 10 sets of 10 or 11 term pairs.		√	√
Nelson et al. [37]	94 undergraduate students rewarded with academic credits	1,016 term pairs selected unsystematically from a cued recall database of 2,000+ pairs.	√		
Janowicz et al. [21]	28 unpaid subjects (20-30 years of age).	Six geographic terms related to bodies of water.		√	√
GeReSiD (see Section 4)	203 unpaid English native speakers.	97 geographic terms from OpenStreetMap; 50 term pairs.	√	√	√

Table 1 Semantic relatedness and similarity gold standards

Furthermore, most datasets do not capture the distinction between semantic similarity and relatedness, and do not analyse the IRA and IRR. It is important to note that most authors did not have the explicit intention to construct gold standards, but rather to analyse specific aspects of semantic similarity or relatedness. However, in some cases, these datasets have been treated as gold standards in the subsequent literature [45, 33]. To the best of our knowledge, only WordSimilarity-353 was explicitly designed to be a generic gold standard.

Some of these gold standards have been extensively utilised to assess general term-to-term similarity measures [45, 33, 13]. In the geographic context, only the MDSM evaluation dataset is suitable to evaluate semantic similarity of geographic terms [44]. However, no existing dataset focusing on geographic terms accounts explicitly for the difference between semantic relatedness and semantic similarity.

4 Geo Relatedness and Similarity Dataset (GeReSiD)

This section presents the Geo Relatedness and Similarity Dataset (GeReSiD), a dataset of human judgements that we have developed to provide a ground truth for the assessment of computational relatedness and similarity measurements. GeReSiD captures explicitly the difference between geo-semantic relatedness and similarity on a sample of geographic terms larger than existing similarity datasets surveyed in Section 3, including both natural and man-made terms. In order to ensure its validity as a gold standard, it focuses on a sample of evenly distributed relatedness/similarity judgements, ranging from very high to very low. Section 4.1 describes our methodology precisely, in order to provide guidelines on constructing datasets to ground the evaluation of measures of geo-semantic relatedness and similarity. Subsequently, Section 4.2 outlines the results obtained from the online survey.

4.1 Survey design

The psychological judgements about geo-semantic relatedness and similarity were collected via an online survey, through an interactive Web interface specifically designed for this purpose. Online surveys constitute a powerful research tool, with well-known advantages and disadvantages [57]. Given the focus of this study on generic terms found in web maps, subjects involved in projects such as OpenStreetMap represent an ideal virtual community of map users and producers to conduct a psychological evaluation. An online survey is an inexpensive and effective way to reach these online communities.

A cross-disciplinary consensus exists on the fact that semantic judgements are affected by the *context* in which the terms are considered [44, 22]. Rodríguez and Egenhofer [44] asked their subjects to rank geographic terms in the following contexts: 'null context,' 'play a sport,' 'compare constructions,' and 'compare transportation systems.' The subjects' attention was therefore focused on

specific aspects of the terms being analysed, rather than on the terms in an unspecified setting.

Although context affects the assessment of semantic similarity, in this survey we aim at capturing the overall difference between semantic relatedness and similarity of terms, without focusing on specific aspects of the conceptualisation. This comparison is an important research topic, frequently mentioned but rarely addressed directly through empirical evaluation. Introducing specific contexts into our survey would increase the complexity of the study by introducing new biases, making the direct comparison between similarity and relatedness problematic. For example, adding a specific context does not increase the inter-subject agreement: in their evaluation, Rodríguez and Egenhofer [44] report a considerably lower association between subjects in the case of context-specific questions (mean Kendall's W being .5), than with a-contextual questions (mean W = .68). Moreover, specific contexts would introduce specific biases, which are beyond the scope of Geo Relatedness and Similarity Dataset (GeReSiD).

As a solution to these issues, we frame the evaluation in the general context of popular *web maps*, in which geographic terms are most frequently visualised and utilised by users. This way, the subjects are induced to use their own conceptualisation of the geographic entities. As happens with semantic judgements, subjectivity inevitably affects the subjects' choices. In this study, subjects are free to choose what properties they consider most relevant to the comparison, and the mean of their ratings quantifies the perceived inter-subject similarity and relatedness of the terms. While the study of the context is beyond the scope of this survey, it certainly represents an important direction for future work.

The geographic terms included in this survey are taken from the OpenStreetMap project. In our previous work, we extracted the lexicon utilised in OpenStreetMap into a machine-readable vocabulary, the OSM Semantic Network [6]. To date, the OSM Semantic Network contains a total of about 4,300 distinct terms, called 'tags' in the project's terminology. From this large set of geographic terms, a suitable sample had to be selected. To be included, a term had to be clearly intelligible, well defined on the OSM Semantic Network, as culturally-unspecific as possible, and present in the actual OpenStreetMap vector map. Following these criteria, we manually selected a set C of 400 terms, including a wide range of natural and man-made entities, such as 'sea,' 'lighthouse,' 'landfill,' 'valley,' and 'glacier.' Using the terms in C, we defined a set P containing all possible pairs of geographic terms $\langle a, b \rangle$ where $a, b \in C$, for a total of 160,000 pairs. We subsequently removed from P symmetric pairs (e.g. removing $\langle b, a \rangle$ when $\langle a, b \rangle$ is defined) and identities (e.g. $\langle a, a \rangle$), resulting in 76,000 valid pairs.

In order to detect issues in the survey, a pilot study was then conducted with 12 graduate students at University College Dublin. A set P_{rand} was constructed by selecting 100 pairs randomly from P. Each pair was associated with a 5-point Likert scale, ranging from low to high relatedness/similarity. The subjects were asked to rate each pair both for semantic relatedness and

similarity, and were then interviewed informally, to obtain direct feedback about the survey. Several useful observations were obtained from this pilot survey. First, most subjects found the test too long. A smaller sample size had to be selected, considering a trade-off between number of pairs and the completion time, in order to ensure that enough subjects would complete the task without losing concentration. Based on the opinion of subjects, we identified 50 pairs as the maximum size of the task, with a completion time of around five minutes, suitable for an unpaid online questionnaire.

In the OpenStreetMap semantic model, tags are made of a key and a value (e.g. *amenity=school*). In the pilot survey, this formalism had to be explained to the subjects, who generally found it confusing. For example, the psychological comparison between *amenity=school* and *amenity=community_centre* was influenced by the shared word 'amenity,' which is highly generic and ambiguous. To make the dataset independent from the peculiar OpenStreetMap tag structure, we extracted short labels for all the 400 terms from the terms' definitions. For example, *amenity=food_court* was labelled as 'food court,' *shop=music* as 'music shop.' In order to increase their semantic clarity, the terms were manually mapped to the corresponding terms in WordNet (see Table 2).

The fully random set of 100 pairs P_{rand} used in the pilot survey obtained a distribution heavily skewed towards low similarity and relatedness. To reach a more uniform distribution, we introduced a partial manual selection in the process. In order to obtain an even distribution in the resulting relatedness and similarity scores, we manually extracted from the pilot survey a set of 50 pairs rated by the 12 subjects as highly related/similar pairs (P_{high}), and 50 middle relatedness/similarity pairs (P_{med}). It is worth noting that while the selection of highly related/similar pairs is intuitive, middle-relatedness/similarity pairs is more challenging, and requires dealing with highly subjective conceptualisations. This aspect is reflected in the survey results (see Section 4.2). The final set of 50 pairs for the questionnaire P_q was assembled from the following elements:

- 16 high-relatedness/similarity pairs (random sample from P_{high})
- 18 middle-relatedness/similarity pairs (random sample from P_{med})
- 16 low-relatedness/similarity pairs (random sample from P)

The pilot survey also showed clearly that assigning both the relatedness *and* the similarity tasks to the same subject was impractical, and was deemed confusing by all subjects who did not possess specific expertise in linguistics. For this reason, we opted to assign randomly only one task to each subject, either on relatedness or similarity, without trying to explain to them the technicalities of this distinction. Instead, we relied on the subjects' inductive understanding of the task through correct examples. Thus, in order to collect reliable judgements on similarity and relatedness, we defined two separate questionnaires, one on relatedness (Q_{REL}), and one on similarity (Q_{SIM}). The two questionnaires were identical, with the exception of the description of the task, and

Term	OpenStreetMap tag	WordNet synset
bay	natural=bay	bay#n#1
sea	place=sea	sea#n#1
basketball court	sport=basketball	basketball court#n#1
beauty parlor	shop=beauty	beauty_parlor#n#1
floodplain	natural=floodplain	floodplain#n#1
greengrocer	shop=greengrocer	greengrocer#n#1
historic castle	historic=castle	castle#n#2
motel	tourism=motel	motel#n#1
political boundary	boundary=political	boundary#n#1
school	amenity=school	school#n#1
stadium	building=stadium	stadium#n#1
...	...	...

Table 2 Sample of terms in GeReSiD. The dataset contains 97 geographic terms.

the labels on the Likert scale (one with a 'dissimilar-similar' scale, the other with 'unrelated-related').

To avoid terminological confusion, the survey was named 'Survey on comparison of geographic terms,' without mentioning either 'similarity' or 'relatedness' in the introductory text. The examples used to illustrate semantic relatedness (*apples - bananas*, *doctor - hospital*, *tree - shade*) and similarity (*apples - bananas*, *doctor - surgeon*, *car - motorcycle*) were based on those by Mohammad and Hirst [34]. A random redirection to either Q_{REL} or Q_{SIM} was then implemented to ensure the random sampling of subjects into two groups, one for similarity and one for relatedness. As the similarity judgement was reported as more difficult than relatedness, we set the probability of a random redirection to Q_{SIM} at $p = .65$, to obtain more responders for similarity. Each subject was only exposed to one of the two questionnaires.

Six general demographic questions about the subject were included: age group, mother tongue, gender, and continent of origin. A textbox was available to type feedback and comments about the survey. The core of each questionnaire was the seventh question, i.e. the relatedness or similarity rating task. The subject had to rate 50 pairs of geographic terms based on their relatedness or similarity, on a 1 to 5 Likert scale. Although the impact of size of the Likert scale, typical options being 5, 7 or 10, is debated in the social sciences, it has little impact on the rating means [11]. If the terms were not clear to the user, a 'no answer' option had to be selected.

Another aspect discussed in the similarity psychological literature is the counterintuitive fact that similarity judgements tend to be asymmetric (e.g. $sim(building, hospital) \neq sim(hospital, building)$) [54]. As this aspect is outside the scope of this study, the order in each pair $\langle a, b \rangle$ was randomised to limit the symmetric bias, i.e. the potential difference between $sim(a, b)$ and $sim(b, a)$ from the subject's perspective. Moreover, a fixed presentation order of pairs can trigger specific semantic associations between terms, and would reduce the quality of the last pairs, rated when the subjects are more likely to be tired. To reduce this sequential-ordering bias, the presentation order of

the pairs was randomised automatically for each subject at the Web interface level.

At the end of the design process, the survey dataset contained 50 pairs of geographic terms to be rated on 5-point Likert scales, including 97 OpenStreetMap terms, with three terms being repeated twice. The pairs were selected to ensure an even distribution between low, medium and high relatedness/similarity. The rating was to be executed in two independent questionnaires, one for semantic similarity (Q_{SIM}) and one for semantic relatedness (Q_{REL}), randomly assigned to the human subjects. In February 2012, the survey was disseminated in OpenStreetMap and geographic information system (GIS)-related forums and mailing lists.

4.2 Survey results

The online questionnaires on relatedness and similarity received 305 responses, 124 for relatedness and 181 for similarity. Given the nature of online surveys, particular attention has to be paid to the agreement between the human subjects, and the detection of unreliable and random answers. In this survey, raters expressed quantitative judgements on geo-semantic relatedness and similarity on a 5-point Likert scale. Two important statistical aspects to be discussed are the interrater reliability (IRR) and the interrater agreement (IRA) [30]. IRR considers the *relative* similarity in ratings provided by multiple raters (i.e. the human subjects) over multiple targets (i.e. the term pairs), focusing on the order of the targets. IRA, on the other hand, captures the *absolute* homogeneity between the ratings, looking at the specific rating chosen by raters.

Several indices have been devised to capture IRR and IRA in psychological surveys [7, 30]. Most indices range between 0 (total disagreement) and 1 (perfect agreement). For example, the ratings of two raters on three targets $\{1, 2, 3\}$ and $\{2, 3, 4\}$ obtain a $IRR = 1$ and $IRA = 0$: the subjects agree perfectly on the ordering of the targets, while disagreeing on all absolute ratings. LeBreton and Senter [30] recommend using several indices for IRR and IRA, to avoid the bias of any single index. We thus include the following indices of IRA and IRR: the mean Pearson's correlation coefficient [43]; Kendall's W [25]; Robinson's A [42]; Finn coefficient [14]; James, Demaree and Wolf's $r_{WG(J)}$ [18].

The 305 responders included both native (208) and non-native English speakers (97). We observed a substantially lower inter-subject agreement when including non-native speakers ($r_{WG(J)} < .5$): the wider variability in these results is due to the varying knowledge of English of these subjects, who might have associated terms to 'false friends' in their native language, i.e. expressions in two different languages that look or sound similar, but differ considerably in meaning. For example, Italian speakers may confuse the meaning of 'factory' with 'farm' ('fattoria' in Italian). Hence, they were excluded from the dataset. Furthermore, three subjects did not complete the task, and their responses were discarded. In order to detect random answers, we computed the

		Relatedness Q_{REL}	Similarity Q_{SIM}	Overall –
Responders	total N	81	122	203
Gender	Male	72	93	165
	Female	9	29	38
Age	18-25	28	39	67
	26-35	14	41	55
	36-45	12	23	35
	46-55	15	10	25
	56-65	7	9	16
	>65	5	-	5
Continent	Africa	-	3	3
	Asia	-	1	1
	Europe	58	95	153
	North America	11	20	31
	South America	-	-	-
	Oceania	12	3	15
Web map expertise	Never used	6	14	20
	Occasional user	18	33	51
	Frequent user	37	39	76
	Expert	20	36	56

Table 3 Demographics of human subjects in GeReSiD

correlation between every individual subject and the means. This way, two subjects in the similarity test showed no correlation at all with the mean ratings (Spearman's $\rho \in [-.05, .05]$), and were removed from the dataset.

Of the resulting dataset, Table 3 summarises demographic information (age group, gender, continent of origin, and self-assessed map expertise). As is possible to observe, the subjects tend be young, male, European, and frequent users of web maps.[5] Table 4 focuses on the indices of IRR and IRA. Following Resnik [41], we consider upper bound on the cognitive plausibility of a computable measure to be the highest correlation obtained by a human rater with the means (e.g. $\rho = .92$ for relatedness). The table shows these upper bounds both for Spearman's ρ and Kendall's τ. All the IRR and IRA indices indicate very similar results, falling in the range $[.61, .67]$. Given the highly subjective nature of semantic conceptualisations, this correlation is satisfactory, and is comparable with analogous psychological surveys [44].

Given the set of term pairs, and the set of human raters, we computed the relatedness/similarity scores as the *mean ratings*, normalised in the interval $[0, 1]$, where 0 means no relatedness/similarity, and 1 maximum relatedness/similarity. As we have stated in the survey objectives, the distribution of such scores should be as even as possible, to ensure that a semantic measure performs well across the board, and not only in a specific region of the semantic relatedness/similarity space. Several pairs in the dataset contain related but not similar terms, and the scores confirm this difference. More specifically, <*sea,island*> obtained a relatedness score of .74 and a similarity of .4. Sim-

[5] Although a better gender, age, and geographic balances would be desirable, we found it difficult to obtain it in practice without drastically limiting the size of the sample.

		Relatedness Q_{REL}	Similarity Q_{SIM}
IRR	mean Pearson's r	.64*	.65*
IRA	Kendall's W	.65*	.64*
	Robinson's A	.62	.61
	Finn coefficient	.65*	.66*
	$r_{WG(J)}$	.66	.67
Upper bound	Spearman's ρ	.92*	.93*
	Kendall's τ	.79*	.82*

Table 4 Indices for interrater reliability (IRR) and interrater agreement (IRA) in GeReSiD. (*) $p < .001$.

Pair			Mean		Agreement	
#	Term A	Term B	REL	SIM	REL	SIM
1	motel	hotel	.93	.90	.86	.82
2	public transport station	railway platform	.87	.81	.80	.72
3	stadium	athletics track	.85	.76	.74	.63
4	theatre	cinema	.82	.87	.57	.79
5	art shop	art gallery	.78	.75	.58	.60
	...	...	...	...	...	...
46	water ski facility	office furniture shop	.05	.05	.92	.88
47	greengrocer	aqueduct	.04	.03	.91	.95
48	interior decoration shop	tomb	.03	.05	.96	.92
49	political boundary	women's clothes shop	.02	.02	.96	.93
50	nursing home	continent	.01	.02	.97	.95

Table 5 Sample term pairs in GeReSiD, with mean geo-semantic relatedness, similarity, and agreement

ilarly, <*mountain hut, mountaintop*> obtained respectively .71 and .49 for relatedness and similarity.

A dimension that has not been addressed in existing similarity gold standards is that of the *pair agreement*, i.e. the consistency of ratings expressed by all subjects on a single pair (see Section 3). For this purpose, we adopt James, Demaree and Wolf's r_{WG}, a popular index to measure IRA on a single target, based on the rating variance [18]. Each pair in Q_{REL} and Q_{SIM} obtains an agreement measure $\in [0, 1]$, where 0 indicates a squared distribution (i.e. raters gave all ratings in equal proportion), and 1 is perfect agreement (i.e. all raters assigned exactly the same rating to the pair). Table 5 shows the content of the resulting dataset, including mean ratings and pair agreement.

Figures 1 and 2 show several statistical characteristics of the resulting dataset, for the 50 pairs in Q_{REL} and Q_{SIM}. Plot 1(a) shows the density of the final relatedness/similarity scores, i.e. the normalised mean rankings in range $[0, 1]$. While the similarity is skewed towards the range $[0, .4]$, the relatedness has slightly more scores in the range $[.4, 1]$, resulting in symmetrical densities. This clearly reflects the fact that semantic similarity is a specific type of semantic relatedness, and semantic similarity is generally lower than relatedness. This can be also observed in the sum of the 50 relatedness scores

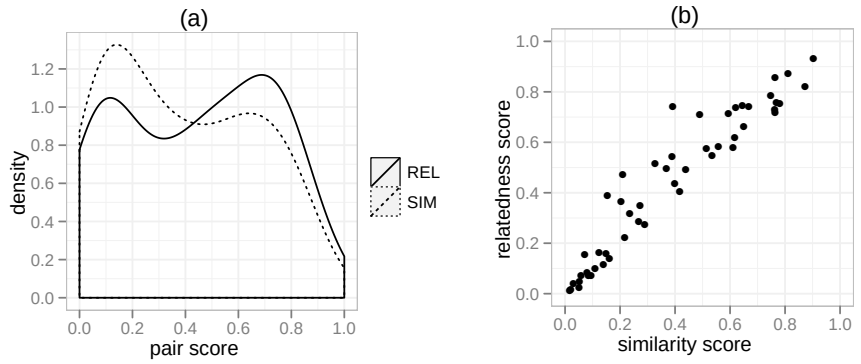

Fig. 1 GeReSiD: REL: semantic relatedness; SIM: semantic similarity. (a) Density of pair score; (b) scatterplot of relatedness versus similarity.

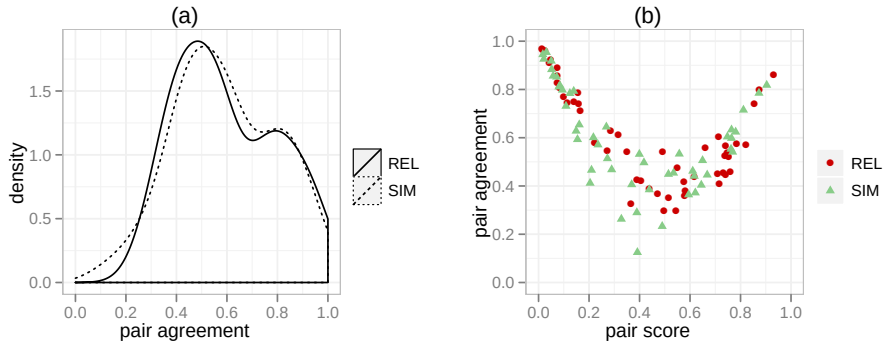

Fig. 2 GeReSiD: REL: semantic relatedness; SIM: semantic similarity. (a) Density of pair agreement; (b) scatterplot of pair agreement and pair score.

($sum = 22.01, mean = .44$) against the similarity scores ($sum = 19.5, mean = .39$). The paired Wilcoxon signed rank test [56] indicates that the relatedness scores are higher than the corresponding similarity ones, at $p < .001$. This trend is clearly visible in plot 1(b). Overall, these densities show that all the score range $[0, 1]$ is satisfactorily covered, i.e. the dataset does not show large gaps.

Plots 2(a) and 2(b) show the properties of pair agreement (index r_{WG}), reporting the relationship between relatedness and similarity, the density of pair agreement, and the relationship between pair agreement and relatedness/similarity scores. In terms of pair agreement, relatedness and similarity follow very close patterns, with a peak $\approx .5$. This agreement might seem low, but it is largely expected, due to the subjective interpretation of the values on the Likert scale.

An explanation of this trend in the pair agreement lies in the fact that humans give consistently different ratings to the same objects: some subjects tend to be strict, and some lenient, resulting in different relative ratings, and therefore low absolute pair agreement [30]. In this regard, a clear pattern emerges from plot 1(b). Pair agreement tends to be high ($> .7$) at the extremes of the scores, when the relatedness/similarity judgement is very low ($[0, .25)$, no relation) or very high ($(.75, 1]$, strong relation). On the other hand, pairs with middle scores (in the interval $[.25, .75]$) tend to have low pair agreement. Relatedness and similarity do not show important differences with respect to pair agreement ($sum = 30, mean = .6$ for relatedness, $sum = 29.6, mean = .59$ for similarity). This detailed analysis, in particular in relation to IRR and IRA, confirms the statistical soundness of GeReSiD, which can be used to assess the cognitive plausibility of computational measures of geo-semantic relatedness and similarity.

5 Conclusions

To date, despite its great potential in GIR and information integration, geo-semantic relatedness has been only marginally studied. In this article, we have discussed a notion of geo-semantic relatedness based on Lehrer's theory of semantic fields, contrasting it with the widely studied geo-semantic similarity. Despite the variety and importance of computational measures devised in natural language processing, the evaluation of such measures remains a difficult and complex task [12].

In order to provide an evaluative baseline for geo-semantic research on relatedness and similarity, we have designed, collected, and analysed the Geo Relatedness and Similarity Dataset (GeReSiD). This dataset contains human judgements about 50 term pairs on semantic relatedness and similarity, covering 97 unique geographic terms. To increase the dataset's usability and clarity, the terms have been mapped to the corresponding terms in WordNet. The judgements were collected from 203 English native speakers, through a randomised online survey. GeReSiD is freely available online, released under an Open Knowledge license.[6] The following points deserve particular consideration:

- The human judgements have interrater agreement (IRA) and interrater reliability (IRR) in the interval $[.61, .67]$. Considering the type of psychological test, this is a fair agreement, indicating that the dataset can be used to evaluate computational measures of semantic relatedness and similarity for geographic terms.
- Human subjects strongly agree on cases of very high and low semantic relationships, and tend to have lower agreement on the intermediate cases.
- Semantic relatedness and similarity are strongly correlated ($\tau = .84, \rho = .95$). Furthermore, semantic relatedness scores are consistently higher than

[6] `http://github.com/ucd-spatial/Datasets` (acc. Apr 10, 2013)

semantic similarity, confirming the more specific nature of semantic similarity.
- The contribution of GeReSiD lies both in its design and validation methodology, as well as the dataset itself. The raw data and the resulting dataset are available for analysis and re-use under a Creative Commons license.
- GeReSiD constitutes an evaluative baseline to evaluate measures of semantic similarity and relatedness. Furthermore, it permits the empirical determination of whether a given measure better approximates similarity or relatedness through the direct comparison of rankings or scores.
- A variety of techniques can be used to compare the rankings or scores generated by a computational measure with GeReSiD, including correlation coefficients (Spearman's ρ and Kendall's τ), and categorical approaches (Cohen's kappa or Fisher's exact test).

Although GeReSiD provides a novel resource to evaluate computational measures of geo-semantic relatedness and similarity, several questions remain open. GeReSiD distinguishes between geo-semantic relatedness and similarity, but not among different *contexts*. As context has been identified as a key aspect of semantic similarity [26], new datasets should be generated to capture explicitly the differences in geo-similarity and relatedness judgements with respect to different contexts, such as appearance and affordances. The investigation of what specific geographic aspects are used by subjects in their judgements also constitutes important future work. The dataset's IRA and IRR are comparable to similar datasets, but have a large margin of improvement.

As Ferrara and Tasso [12] point out, this evaluative approach has several limitations. Human subjects understand intuitively semantic relatedness and similarity, but the translation of such judgements into a number is very subjective. Different information communities can express different judgements on the same term pairs. Alternative approaches to the evaluation of computational measures should be investigated, aiming at cross-validating the findings generated by GeReSiD. A promising route might consist of evaluating human-readable *explanations* of relatedness measures, and not only numeric scores or rankings [17]. Moreover, the collection of judgements was conducted through online surveys in an uncontrolled environment, which have well-known issues [57].

Ultimately, the cognitive plausibility is assessed using correlation indexes such as Spearman's ρ and Kendall's τ, which have specific limitations. For example, they tend to attribute the same weight to high and low similarity rankings, whilst computational applications normally need more precision on highly-related/similar pairs, which tend to be utilised in GIR and information integration. Using GeReSiD as input data, new techniques to assess cognitive plausibility can be developed, offering tools tailored to the study of geo-semantic relatedness and similarity. Fruitful future work, as geo-semantic similarity is a specific case of geo-semantic relatedness, will consist of the generalisation of existing geo-similarity theories to the framework of geo-semantic relatedness.

Acknowledgements The research presented in this article was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

References

1. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19–27), ACL.
2. Bakillah, M., Bédard, Y., Mostafavi, M., Brodeur, J. (2009). SIM-NET: A View-Based Semantic Similarity Model for Ad Hoc Networks of Geospatial Databases. *Transactions in GIS*, *13*(5-6), 417–447.
3. Ballatore, A., Wilson, D., Bertolotto, M. (2012). The Similarity Jury: Combining expert judgements on geographic concepts. In: Castano, S., Vassiliadis, P., Lakshmanan, L., Lee, M. (Eds) *Advances in Conceptual Modeling. ER 2012 Workshops (SeCoGIS)*, LNCS, vol. 7518, Springer, pp. 231–240.
4. Ballatore, A., Bertolotto, M., Wilson, D. (2013). Computing the Semantic Similarity of Geographic Terms Using Volunteered Lexical Definitions. *International Journal of Geographical Information Science*, *27*(10), 2099–2118.
5. Ballatore, A., Bertolotto, M., Wilson, D. (2013). Geographic Knowledge Extraction and Semantic Similarity in OpenStreetMap. *Knowledge and Information Systems*, *37*(1), 61–81.
6. Ballatore, A., Wilson, D., Bertolotto, M. (2013). A Survey of Volunteered Open Geo-Knowledge Bases in the Semantic Web. In: Pasi, G., Bordogna, G., Jain, L. (Eds) *Quality Issues in the Management of Web Information*, Intelligent Systems Reference Library, vol. 50, Springer, pp. 93–120.
7. Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D. (1999). Beyond Kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*(1), 3–23.
8. Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.
9. Budanitsky, A., Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, *32*(1), 13–47.
10. Cimiano, P., Völker, J. (2005). Towards large-scale, open-domain and ontology-based named entity classification. In: *Recent Advances in Natural Language Processing, RANLP 2005* (pp. 166–172), ACL.
11. Dawes, J. (2008). Do data characteristics change according to the number of scale points used? *International Journal of Market Research*, *50*(1), 61–78.
12. Ferrara, F., Tasso, C. (2013). Evaluating the Results of Methods for Computing Semantic Relatedness. In: Gelbukh, A. (ed.) *Computational Lin-*

guistics and Intelligent Text Processing, LNCS, vol. 7816, Springer, pp. 447–458.
13. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E. (2002). Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, *20*(1), 116–131.
14. Finn, R. (1970). A Note on Estimating the Reliability of Categorical Data. *Educational and Psychological Measurement*, *30*(1), 71–76.
15. Goldstone, R., Son, J. (2005). Similarity. In: Holyoak, K., Morrison, R. (Eds) *Cambridge Handbook of Thinking and Reasoning*, New York: Cambridge University Press, pp. 13–36.
16. Hecht, B., Raubal, M. (2008). GeoSR: Geographically Explore Semantic Relations in World Knowledge. In: *The European Information Society: Taking Geoinformation Science One Step Further*, LNGC, Springer, pp. 95–113.
17. Hecht, B., Carton, S. H., Quaderi, M., Schöning, J., Raubal, M., Gergle, D., Downey, D. (2012). Explanatory semantic relatedness and explicit spatialization for exploratory search. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 415–424), ACM.
18. James, L., Demaree, R., Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of applied psychology*, *69*(1), 85–98.
19. Janowicz, K., Raubal, M. (2007). Affordance-based similarity measurement for entity types. In: *Spatial Information Theory*, LNCS, vol. 4736, Springer, pp. 133–151.
20. Janowicz, K., Keßler, C., Schwarz, M., Wilkes, M., Panov, I., Espeter, M., Bäumer, B. (2007). Algorithm, implementation and application of the SIM-DL similarity server. In: *GeoSpatial Semantics: Second International Conference, GeoS 2007* (pp. 128–145), Springer, LNCS, vol. 4853.
21. Janowicz, K., Keßler, C., Panov, I., Wilkes, M., Espeter, M., Schwarz, M. (2008). A study on the cognitive plausibility of SIM-DL similarity rankings for geographic feature types. In: Fabrikant, S., Wachowicz, M. (Eds) *The European Information Society: Taking Geoinformation Science One Step Further*, LNGC, Springer, pp. 115–134.
22. Janowicz, K., Raubal, M., Schwering, A., Kuhn, W. (2008). Semantic Similarity Measurement and Geospatial Applications. *Transactions in GIS*, *12*(6), 651–659.
23. Janowicz, K., Raubal, M., Kuhn, W. (2011). The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, *2*(1), 29–57.
24. Kaptchuk, T. (2001). The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology*, *54*(6), 541–549.
25. Kendall, M., Smith, B. (1939). The problem of m rankings. *The annals of mathematical statistics*, *10*(3), 275–287.

26. Keßler, C. (2007). Similarity measurement in context. In: *Proceedings of the 6th International and Interdisciplinary Conference on Modeling and Using Context* (pp. 277–290), Springer, LNCS, vol. 4635.
27. Keßler, C. (2011). What is the difference? A cognitive dissimilarity measure for information retrieval result sets. *Knowledge and Information Systems*, *30*(2), 319–340.
28. Khoo, C., Na, J. (2006). Semantic relations in information science. *Annual Review of Information Science and Technology*, *40*(1), 157–207.
29. Kuhn, W. (2013). Cognitive and linguistic ideas and geographic information semantics. In: *Cognitive and Linguistic Aspects of Geographic Space*, LNGC, Springer, pp. 159–174.
30. LeBreton, J., Senter, J. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852.
31. Lehrer, A. (1985). The influence of semantic fields on semantic change. In: Fisiak, J. (ed.) *Historical Word Formation*, Berlin: Walter de Gruyter, pp. 283–296.
32. Medin, D., Goldstone, R., Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science*, *1*(1), 64–69.
33. Miller, G., Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28.
34. Mohammad, S., Hirst, G. (2012). Distributional Measures of Semantic Distance: A Survey. *Computing Research Repository (CoRR)*, *abs/1203.1858*, 1–39, URL `http://arxiv.org/abs/1203.1858`.
35. Montello, D. R., Fabrikant, S. I., Ruocco, M., Middleton, R. S. (2003). Testing the first law of cognitive geography on point-display spatializations. In: Kuhn, W., Worboys, M., Timpf, S. (Eds) *Spatial Information Theory. Foundations of Geographic Information Science*, LNCS, vol. 2825, Springer, pp. 316–331.
36. Morris, J., Hirst, G. (2004). Non-classical lexical semantic relations. In: *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics* (pp. 46–51), ACL.
37. Nelson, D., Dyrdal, G., Goodmon, L. (2005). What is preexisting strength? Predicting free association probabilities, similarity ratings, and cued recall probabilities. *Psychonomic Bulletin & Review*, *12*(4), 711–719.
38. Pedersen, T., Kolhatkar, V. (2009). Wordnet::senserelate::allwords: a broad coverage word sense tagger that maximizes semantic relatedness. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session* (pp. 17–20), ACL.
39. Pedersen, T., Patwardhan, S., Michelizzi, J. (2004). WordNet::Similarity: measuring the relatedness of concepts. In: *Proceedings of Human Language Technologies: The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume:*

Demonstration Session (pp. 38–41), ACL.
40. Rada, R., Mili, H., Bicknell, E., Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, *19*(1), 17–30.
41. Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI'95* (pp. 448–453), Morgan Kaufmann, vol. 1.
42. Robinson, W. (1957). The statistical measurement of agreement. *American Sociological Review*, *22*(1), 17–25.
43. Rodgers, J., Nicewander, W. (1988). Thirteen ways to look at the correlation coefficient. *American Statistician*, *42*(1), 59–66.
44. Rodríguez, M., Egenhofer, M. (2004). Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure. *International Journal of Geographical Information Science*, *18*(3), 229–256.
45. Rubenstein, H., Goodenough, J. (1965). Contextual Correlates of Synonymy. *Communications of the ACM*, *8*(10), 627–633.
46. Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics*, *24*(1), 97–123.
47. Schwering, A. (2008). Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. *Transactions in GIS*, *12*(1), 5–29.
48. Schwering, A., Kuhn, W. (2009). A hybrid semantic similarity measure for spatial information retrieval. *Spatial Cognition & Computation*, *9*(1), 30–63.
49. Schwering, A., Raubal, M. (2005). Spatial relations for semantic similarity measurement. In: *Perspectives in Conceptual Modeling* (pp. 259–269), Springer, LNCS, vol. 3770.
50. Strube, G. (1992). The role of cognitive science in knowledge engineering. *Contemporary knowledge engineering and cognition*, *622*, 159–174.
51. Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. In: *Economic geography. Supplement: Proceedings. International Geographical Union. Commission on Quantitative Methods*, vol. 46, Worcester, MA: Clark University, pp. 234–240.
52. Toutanova, K., Klein, D., Manning, C., Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (pp. 173–180), ACL, vol. 1.
53. Turney, P. (2006). Similarity of semantic relations. *Computational Linguistics*, *32*(3), 379–416.
54. Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327–352.
55. Wang, C., Wang, J., Xie, X., Ma, W. Y. (2007). Mining geographic knowledge using location aware topic model. In: *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval* (pp. 65–70), ACM.

56. Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.
57. Wright, K. (2005). Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication*, *10*(3), URL `http://jcmc.indiana.edu/vol10/issue3/wright.html`, article 11.