Improving Comprehension of Measurements Using Concrete Re-Expression Strategies

Jessica Hullman[1], **Yea-Seul Kim**[1], **Francis Nguyen**[1], **Lauren Speers**[2], **and Maneesh Agrawala**[3]

[1]University of Washington
Seattle, WA, USA
jhullman,yeaseul1,fmnguyen@uw.edu

[2]University of California
Berkeley, CA, USA
lspeers@berkeley.edu

[3]Stanford University
Palo Alto, CA, USA
maneesh@cs.stanford.edu

ABSTRACT

It can be difficult to understand physical measurements (e.g., 28 lb, 600 gallons) that appear in news stories, data reports, and other documents. We develop tools that automatically re-express unfamiliar measurements using the measurements of familiar objects. Our work makes three contributions: (1) we identify effectiveness criteria for objects used in concrete measurement re-expressions; (2) we operationalize these criteria in a scalable method for mining a large dataset of concrete familiar objects with their physical dimensions from Amazon and Wikipedia; and (3) we develop automated concrete re-expression tools that implement three common re-expression strategies (adding familiar context, reunitization and proportional analogy) as energy minimization algorithms. Crowdsourced evaluations of our tools indicate that people find news articles with re-expressions more helpful and re-expressions help them to better estimate new measurements.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Measurement re-expression, analogy, reunitization, proportional analogy.

INTRODUCTION

People often encounter measurements of physical properties (e.g., height, length, weight, volume) in daily life, such as when reading news stories or data reports, interacting with visualizations, or buying household products online. Measurements are often difficult for people to understand because people tend to have limited experience with thinking about physical measurements explicitly in daily life [32, 35]. Uncommonly large (or small) magnitudes (e.g. 4 tons) and unfamiliar units (e.g. volumetric units like ft^3 or m^3) can exacerbate such difficulties. Misunderstanding measurements can lead to major consequences; for example if news readers cannot grasp how much California water the average American consumes by purchasing fruits grown in the state (300 gallons)[1], or how high storm surge flood waters from a hurricane could rise in their area (4 ft)[2], they may not be compelled to change their grocery list to help conserve water, or evacuate their home.

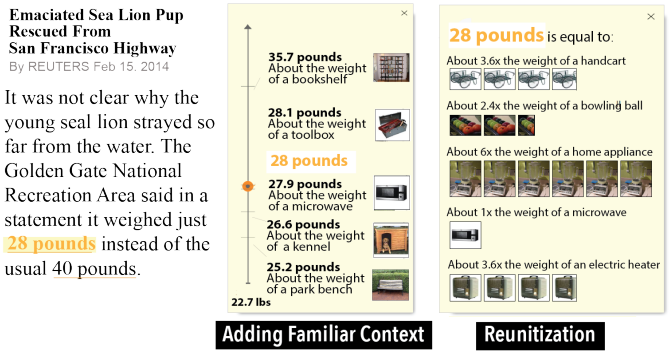

Figure 1. A text article with our automated concrete re-expression tools using two common strategies: adding familiar context (left) and reunitization (right) to provide more context for the measurements by comparing them to measurements of familiar objects.

One common technique for helping people make sense of a new measurement is to compare it to the measurement of a familiar concrete object [25, 23, 24, 28, 29, 37]. Studies indicate that familiar reference objects are often similar even across individuals (e.g. when thinking of hand-held objects many think of a golf or tennis ball) [37, 25, 23, 24]. Various forms of such *concrete re-expressions* take advantage of how people often think of measurements in relation to the measurements of objects they are familiar with to help them reason about measurements more accurately:

- ***Adding familiar context*** presents a measurement (e.g., 28 lbs) alongside objects with similar measurements (e.g., the weight of a tool box, the weight of a microwave) (Fig. 1 left).
- ***Reunitization*** re-expresses a measurement (e.g., 3 ft) using a more familiar object as the unit with a multiplicative scale factor for converting from one to the other (e.g., 2 times the height of a single bed) (see also Fig. 1 right).
- ***Proportional analogy*** re-expresses a pair of measurements (e.g., the ratio between the volumes of Mercury and Earth) using two familiar objects that have measurements with the same ratio (e.g., the ratio between the volume of a sugar bowl and a watering can) (see also Fig. 2).

[1]https://goo.gl/f6dc7W
[2]https://goo.gl/VGu2oa

CHI 2018, April 21–26, 2018, Montréal, QC, Canada.

ACM ISBN 978-1-4503-5620-6/18/04 $15.00.
http://dx.doi.org/10.1145/3173574.3173608

Re-expressions are widely used in educational settings to introduce students to scales of measurement [25, 28, 29, 37]. For example, proportional analogies help people develop STEM literacy by supporting reasoning at multiple scales, which is common in the sciences [22]. To enhance scientific communication, journalists are instructed to craft re-expressions whenever a measurement might be unfamiliar.[3] [4].

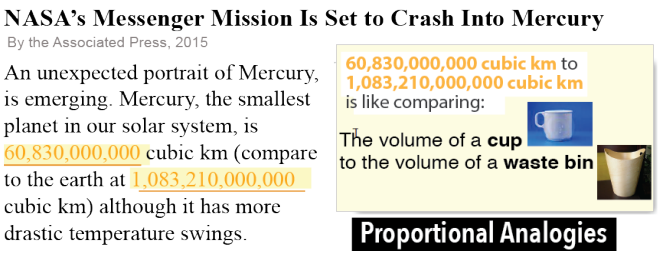

Figure 2. Our re-expression tool provides a *proportional analogy* that compares the unfamiliar measurements in a text article to measurements of familiar objects.

Unfortunately, the manual effort it currently requires for educators, journalists, or designers to create effective measurement re-expressions limits their use in the many settings where measurements appear. Authors must choose one or more familiar objects (considering factors like how close its measurement is to the unfamiliar measurement, how much its size varies across instances, etc.), find reliable measurements for the objects, calculate the conversion, and visualize the re-expression.

Our work contributes tools that make it possible for more people to benefit from concrete measurement re-expressions. Our first contribution is **a set of effectiveness criteria for concrete re-expression objects.** These criteria formalize attributes like the object's familiarity, concreteness, and countability; the familiarity of the object's measurement; and the variance in the measurement across object instances. We next show how these criteria can be operationalized in a **scalable approach to mining object datasets**, including Amazon, DBPedia, and Freebase databases. Our approach applies a three-stage pipeline for constructing a database of familiar objects and their measurements, using semantic databases (WordNet and ImageNet), and crowdsourcing techniques, resulting in a database containing an average of 11.3 instances of 1,192 familiar objects with four measurements (weight, height, length, and volume) for each instance. We contribute a set of **automated concrete re-expression tools** that implement the three re-expression strategies (adding familiar context, re-unitization and proportional analogy) as energy minimization algorithms.

As a proof-of-concept for how measurement re-expressions can be incorporated in an online reading context, we develop Web-based applications that analyze text news articles for measurements and use our automated tools to re-express them using text and visualizations (Fig. 1). Through user studies we show that 1) viewers who saw concrete re-expressions in text articles rated the content as approximately 2 ± 0.5 points more helpful for understanding measurements on a 7-point scale compared to no re-expressions, and that 2) viewers who saw our Adding Familiar Context re-expressions as they estimated unfamiliar measurements were $8\pm1.6\%$ more accurate compared to not seeing re-expressions.

[3]https://goo.gl/aio143
[4]https://goo.gl/KjsDDV

RELATED WORK

Chevalier et al. [12] use an analysis of over 300 examples of visual re-expressions of measurements found in infographic-style visualizations to identify common re-expression strategies like reunitization and proportional analogy. Noting that many semantic and contextual considerations go into selecting a re-expression object, they suggest that it is not possible to develop a general technique for generating re-expressions automatically [13]. We propose that automatic re-expression is possible using a database of familiar objects and we show that the resulting re-expressions are helpful to users.

Closer to our work, Kim et al. [26] developed a method for automatically generating re-expressions of spatial distances and areas using a database of landmarks. Their approach employs an objective function that considers the overall familiarity of a landmark, the proximity of the user to the landmark (representing personal familiarity), and the multiplicative factor to generate reunitizations. We similarly use an objective function to generate measurement re-expressions, but focus on criteria and tools to help people understand on other difficult to understand physical measurements like weight, height, and volume. We contribute a scalable pipeline for creating a database of familiar objects and implementations of two additional strategies: Adding Familiar Context and Proportional Analogy.

Chaganty and Liang [10] automatically produce re-expressions for measurements that include weight, length, and volume using natural language generation techniques applied to a small set of web-scraped statistics. However, their database contains mostly counts of people and money, which they attribute to their reliance on a news corpus for the measurement data. Their approach also does not consider the familiarity of the statistics, resulting in re-expressions that reuse a small set of potentially unfamiliar measures (e.g., the population of Texas, the number of Google employees). Our approach is guided by a set of criteria that we develop to describe the properties objects should have to serve effectively in measurement re-expressions, including familiarity.

While not pursuing a fully automated solution, Barrio et al. [3] present three crowdsourced user studies of re-expressions of measurements generated by crowdworkers. They find that users who view re-expressions can more accurately recall and estimate unfamiliar measurements by roughly 10 to 15% compared to users who did not see re-expressions. To evaluate our automated re-expression algorithms, we adapt their estimation study design. More recently, Riederer et al [33] studied how properties of spatial re-expressions impact their effectiveness for helping users estimate measurements, finding that long term benefits of re-expressions persisted 6 weeks later.

The lack of a large repository of object measurements has also motivated researchers to extract numerical attributes of objects from Web text [19] or text and images [2], including physical measurements like heights and lengths. The goal of such work is to develop automated techniques that use inference

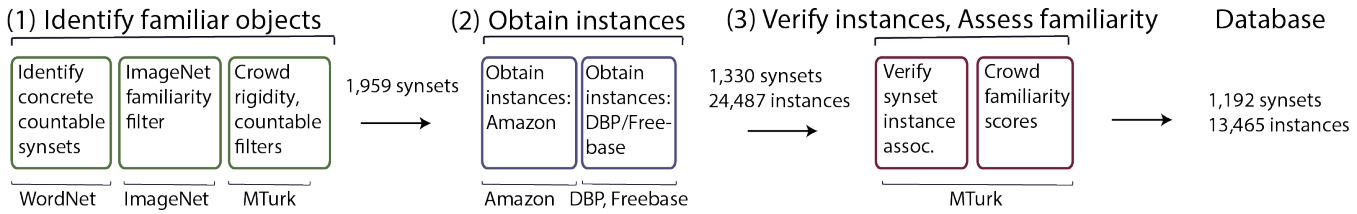

Figure 3. Our three-stage database construction pipeline. In stage 1, we identify candidate classes of objects (represented by WordNet concepts called synsets) that are concrete, countable, rigid and familiar using a combination of WordNet, ImageNet and Amazon Mechanical Turk. In stage 2, we obtain specific instances of the classes of objects (synsets) and their measurements for each synset extracted in stage 1. In stage 3, we use crowdsourcing to verify that instances are associated correctly with the synset and to obtain object and measure familiarity scores.

based on relative size information to estimate measurements that are not directly observed. This results in low accuracy (e.g., 64-73% [19]) overall for inferred absolute measurements. Higher accuracy (83%) is achieved for relative measurements (e.g., is an elephant or a butterfly bigger), but this limits applications [2]. Our primary focus is on collecting accurate measurement data, since it will be presented to end-users to help them understand measurements.

In addition to researchers seeking to address the need for more scalable solutions, several developers have created small hand-curated object databases with corresponding reunitization tools to make it easier for readers of internet documents, journalists, and educators to access measurement context [14, 15]. However, these solutions are limited in scope (e.g., reuse the same objects repeatedly or cannot return a re-expression for some magnitudes), and do not generalize due to their reliance on the designer's intuition. Wolfram Alpha (WA) [39] can also reunitize some input measurements for height, length, weight, and volume, but the size of the database and methods for generating re-expressions are not documented. In contrast we contribute a three-stage pipeline for constructing a large-scale database of familiar objects and their measurements. We provide automated implementations of three re-expression strategies, which we evaluate with several user studies.

CRITERIA FOR EFFECTIVE RE-EXPRESSION OBJECTS

Psychologists, educators and researchers have suggested several criteria for an object to serve as an effective candidate for measurement re-expression [6, 12, 20, 25, 28, 29]. We distill these suggestions and our observations from analyzing published re-expression examples into effectiveness criteria:

- ***Concrete***: The object should have a physical form and measurable properties (e.g. weight, height, volume) so that it can be used to re-express physical measurements.
- ***Countable***: The quantity of the object should be clear so that it can be used as a countable unit.
- ***Rigid***: The object should be rigid and resist bending and folding because such deformations can impact its measurements (e.g., height, volume).
- ***Object Familiarity***: The object should be familiar so that when it is used for re-expression people can immediately relate to it from experience.
- ***Measure Familiarity***: The measured property of the object (e.g. weight, height, volume, etc.), should also be familiar so that when it is used for re-expression people can relate to it from experience.
- ***Low Measure Variance***: The measured property of the object should not vary greatly between instances of the object so that people can reliably estimate the measurement.
- ***Measure Closeness***: The measured property of the object should be close to the unfamiliar input measurement so that people can mentally convert from the measurement of the familiar object to the unfamiliar measurement.
- ***Object Similarity***: For proportional analogies, the two objects should be conceptually similar to one another (e.g. both are toys, both are furniture, etc.) so that they are easier to compare.

A necktie, for example, fulfills all of the re-expression criteria except rigidity: since the length and volume of a necktie may vary depending on how it is folded, it is a poor choice for the re-expression. Pavement is concrete, rigid, and familiar, but is not countable, and its physical measures (e.g. height, length, weight) are not well defined. Alternatively, a crayon fulfills all of the criteria for re-expressing 55 inches except measure closeness. The resulting re-expression – 55 inches is about the length of 46 crayons – equates a large number of crayons with the input measurement, requiring more mental effort to process the conversion. In contrast, a park bench (40 inches) and a writing desk (42 inches) are objects that fulfill all the criteria and result in effective re-expressions of 55 inches.

DATABASE CONSTRUCTION

We present a three stage pipeline for constructing a database of objects covering four common measures (weight, height, length, volume) at multiple scales (e.g. 0.01 lbs to 1000 lbs) (depicted in Fig. 3). In stage 1 we identify familiar classes of objects (e.g., basketball, laptop) using WordNet, ImageNet and crowdsourcing techniques. In stage 2, we collect instances of these objects (e.g. Spaulding NBA Street Basketball, Apple iBook 12.1-Inch Laptop), including their measurements, from Amazon and two sources of Wikipedia data, DBpedia, and Freebase. In stage 3 we use crowdsourcing to filter out instances that crowdworkers cannot easily verify as examples of the intended object and to obtain familiarity information for the objects. Our pipeline is designed to only include objects that are concrete, countable, rigid and familiar, while obtaining enough information for our automated re-expression tools to compute measure familiarity,variance and closeness.

Stage 1: Identify Familiar Classes of Objects (Synsets)

Our goal in the first stage of the pipeline is to identify objects that achieve the first four effectiveness criteria. We focus on these four criteria because we can assess them without access

to specific instances and measurements, and thereby eliminate entire classes of objects from further consideration.

We first consider WordNet [30], a semantic database structured as a graph of distinct concepts called *synsets* that range from objects (e.g., *camera; photographic camera*) to abstract concepts (e.g., *idea; thought*). WordNet represents a synset as a set of synonymous phrases sepeared by semicolons. The WordNet graph is a DAG in which each synset is a node and a *hypernym* edge represents a more general synset, while a *hyponym* edge represents a more specific synset. For example, the synset *digital camera* has the hypernym synset *camera; photographic camera* and the hyponym synset *webcam*.

Identify concrete, countable synsets. All concepts in WordNet originate from the highest level synset *entity*, which has hyponym synsets including *abstraction* and *physical entity*. The synset *physical object* is a hyponym of *physical entity*, and is differentiated into parts of objects (*part; portion*) and whole objects (*whole; unit*), the latter of which has hyponyms including *natural object* and *artifact*. The synset *artifact* is a good proxy for concrete, countable objects since it is defined by WordNet as "a man-made object that can be considered as a single whole." On the other hand, other hyponym synsets of *physical object*, such as *natural object*, include hyponyms that are not necessarily concrete countable objects, such as *universe* or *tangle*. We therefore extract all synsets in WordNet that lie in the *artifact* subtree resulting in 12,011 synsets.

ImageNet familiarity filter. To ensure that the 12,011 artifact synsets represent familiar objects, we look them up in ImageNet [18], an image database organized using the concept graph of WordNet, but with a set of images also representing each synset. ImageNet also includes a popularity percentile score for each synset which is calculated from the number of Google search results returned for the phrase representation of the synset and the frequency of its occurrence in the British National Corpus [8]. We experimentally found that filtering the synsets to retain only those with a popularity score of greater than or equal to 70% eliminated synsets representing specific, less well-known objects (e.g., *armilla*: 30%, a type of *bracelet*: 70%, *Hoover*: 67%, a type of *vacuum cleaner*: 72%). This filter reduces the number of synsets to 2,458.

Crowd rigidity and countability filters. The remaining synsets may include objects that are not rigid (e.g. *miniskirt*, *folding chair*) and that are not easily countable (e.g. *camouflage*). We therefore ask Amazon Mechanical Turk (MTurk) workers to label whether each synset represents a rigid and a countable object. To make our approach to gathering instances of the synsets from Amazon, DBPedia, and Freebase in stage 2 more efficient, we also ask workers whether they think examples of the synset are sold on Amazon. To convey the synset we present the synset phrases and three representative ImageNet images of the synset. (Fig. 4). Each human intelligence task (HIT) asks a worker to label ten synsets, where one of the ten is from a gold-standard set of 130 synsets that the first author labeled a priori. The HIT carries a reward of $0.15.

Figure 4. Representation of a synset for crowdworkers.

Fifty workers participated in the HIT. We obtained 10 responses for each synset, and omitted responses if the worker incorrectly labeled the gold synset in the HIT (omitting 14% of responses). We set the final rigidity and countability and purchasability label for each synset as the majority response. Filtering out the non-rigid and non-countable synsets leaves 1,959 synsets.

Stage 2: Obtain Instances

In the second stage of the pipeline we obtain instances of the synsets (e.g., examples of specific bike models on Amazon for the synset *mountain bike, all-terrain bike*) and their measurements (weight, height, length, volume) for the 1,959 synsets we identify in the first stage. We query Amazon for instances of the 991 synsets that crowdworkers indicated are purchasable on Amazon. In contrast to the specific examples of consumer products sold on Amazon, Wikipedia articles tend to describe either classes of objects in the abstract, similar to WordNet synsets (e.g., *telephoto lens* or *soccer ball*), or specific instances of classes of objects that are not on Amazon (e.g., locations like *Wrigley Field*, historical objects like *Rosetta Stone*, etc.). Because of the different coverage of Wikipedia, we query DBpedia and Freebase for the remaining 968 synsets. DBpedia [27] is a structured, downloadable extract of information from Wikipedia, while Freebase [5] is also a Wikipedia extract but contains some user-generated content.

Obtain instances from Amazon. To extract instances and their measurements from Amazon, we conduct an ItemSearch with each synset phrase using the Product Affiliate API [1]. We found that while the first 20 instances (products) returned by the search are usually representative of the synset, the remaining results may not be as representative (e.g., a search on the synset *basketball* begins to show basketball accessories after about the first 20 results).

For each of the first 20 instances, we directly extract the weight, height, and length from the ItemAttributes description. We compute the volume of the instance's bounding box using the product dimensions, which are reported as triplets of values representing the length, height, and depth of the product.

Some Amazon products are sold in aggregate (e.g. a box of pens). If the count of items for the instance is provided as an ItemAttribute, we also record this count.

For some aggregated products the item count is not listed in the ItemAttributes, but is given in the product title (e.g., "Set of 3 Pens", "Box of 10 Razors"). We use a set of pre-defined regular expressions (e.g., "set of", "box of", "XX-count") to extract the quantity in the product title. If no quantity is obtained, we assume the quantity is one.

When the item count is greater than one, we divide the weight and volume of the instance by the count to produce single unit measures. However, dividing heights and lengths by quantity does not always reflect the the true measure per unit of the instance – for a box of 500 drinking straws the length of the box is usually about the length of each straw. Therefore, we do not adjust heights and lengths for aggregates and instead

eliminate these instances for consideration in all applications that re-express heights or lengths.

This process yields at least one instance for each of the 991 synsets we query for on Amazon.

Obtain instances from DBPedia/Freebase. To include instances of synsets that are not sold on Amazon (e.g., aircraft carrier) in our database, we leverage the structured information available in Wikipedia article *infoboxes*. We access the infobox entries through the DBPedia [27] and Freebase [5] APIs [16, 21]. As with Amazon, we first search for Wikipedia articles using each synset phrase, checking for matches against the article title or article category. For each such matching article (instance), we check if an infobox exists and then extract the measurements (if any) using DBPedia and Freebase.

This process yields at least one instance for 339 of the 968 synsets we query for in DBPedia/Freebase.

Accuracy of Extracted Measurements

We manually checked the accuracy the measurements of 50 instances (39 from Amazon and 11 from DBPedia/Freebase) we extracted and achieved 85% for height, 91% for length, 98% for weight, and 96% for volume. For those instances where the measurement was wrong, the most common causes were either a misreported measurement or measurement unit as returned by the source API (Amazon, DBpedia, or Freebase).

We separately checked the accuracy of our item counts extraction for 50 instances associated with 10 synsets that we deemed likely to be sold in aggregate (e.g., *golf ball*, *dish*). We extracted the correct count for 86% of these instances.

At the end of stage two we have 1,330 synsets for which we have at least one instance and an average of 18.4 instances for each synset giving us a total of 24,487 instances.

Stage 3: Crowd Verification and Familiarity Scores

The final stage of our pipeline uses crowdsourcing to verify that the instances match their associated synsets and to obtain object and measurement familiarity scores for the synsets.

Crowdsourced verification of synset-instance associations. Our automated search-based methods in stage 2 may sometimes return instances that are not representative of the query synset (e.g., the instance "Giro Feature Mountain Bike Helmet" for an Amazon query using the synset *mountain bike, all-terrain bike*). We therefore ask five human judges, recruited from MTurk to check whether each instance obtained in stage 2 correctly represents the associated synset. We present the synset representation (Fig. 4) and a screenshot of the Amazon entry for the instance. Each HIT carries a reward of $0.10 and includes 10 (synset, instance) pairs. In each HIT, one of the pairs is randomly drawn from a gold standard set containing 300 instances which the first author verified manually a priori.

Forty-seven workers completed the task. We omit responses from HITs where the worker mislabeled the gold standard instance (9% of total responses). For each synset, we retain all instances for which the majority of the remaining workers agreed that the instance was an example of the synset. This process filters out a large number of possible instances across the set of synsets: nearly 45% of the instances collected in stage 2 are removed in stage 3. Of the 1,330 synsets at the end of stage 2 we are left with 1,192 synsets with at least one instance and an average of 11.3 instances per synset.

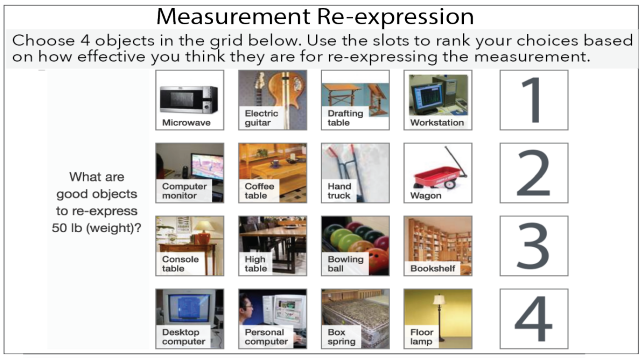

Figure 5. Interface for MTurk HIT in which crowdworkers choose 4 objects from a grid of possible objects (right) that they believe would be effective for re-expressing a reference measurement (left).

Crowdsourced familiarity scores. While the ImageNet popularity percentile scores we gather in stage 1 provide a coarse means of eliminating very rare objects, they do not provide a fine-grained measure of object familiarity (e.g., *flare* with an ImageNet popularity score of 73% is not clearly more familiar than *bracelet* at 70%). The scores also do not address measurement familiarity. Because the magnitude of the measurement being considered for re-expression will influence which objects are perceived as most familiar (e.g., a straight pin might be a familiar object for re-expressing 0.01 lbs but not for expressing 50 lbs), we design a crowdsourced task that asks workers to consider how familiar objects are for re-expressing for each of a set of input measurements that range in magnitude ([0.01, 0.05, 0.1, 0.5, ... 10,000]).

To elicit the familiarity judgments, we use a grid based comparison task as in Wilber et al. [38]). In such tasks, a probe object is presented next to a grid of n other objects (we use n=16), and subjects are asked to select k objects from the grid that have some property (we use k=4). We presented sample measurements as probes and asked crowdworkers to select k familiar objects from the grid that could be used for re-expressing the reference measurement (Fig. 5). For each measure (weight, height, length, and volume) and reference measurement from the set [0.01, 0.05, 0.1, 0.5, ... 10,000], we generated grids of random samples of candidate objects whose measurements were within a pre-defined range of the reference measurement. Specifically we included all synsets with measurements within a multiplicative factor of 0.1 and 50, under the assumption that people would not choose objects with magnitudes very different from the reference.

We generated 515 grids of n=16 synsets. We presented each as a HIT with a reward of $0.05 to 15 workers in the Amazon Mechanical Turk Master pool. 66 workers participated.

We combined the crowd rankings to create a *crowd familiarity score* for each synset, measure, and magnitude. For each worker and reference measurement, we interpret the first synset the worker chose as having rank 1, the second as having rank 2, etc. To penalize objects that did not appear in a grid or

appeared and were not ranked by a worker for a reference measurement, we assign a penalty rank (which we set to 20). Thus for each synset and each reference measurement [0.01, 0.05, ...] we obtain a crowd familiarity score which is the mean rank of the synset for that reference across all workers. This score provides a more nuanced representation of familiarity than the ImageNet-based familiarity score. For example, the ImageNet familiarity score would suggest that the most familiar objects regardless of input measurement are a building or a couch. However, the crowd familiarity score will predict a coffee mug as highly familiar for re-expressing 0.25 gal, but a bathtub as better for re-expressing 300 gal.

The resulting database contains 1,192 unique synsets, with an average of 11.3 instances per synset (total set of instances: 13,465, including 11,405 Amazon, 1,183 DBpedia, and 877 Freebase instances). For each of the 1,192 synsets, we compute the mean and standard deviation for each measure (weight, height, length, and volume) across the crowdworker verified instances of that synset. We also record the mean familiarity rank for each synset. We provide the data as supplemental material, along with APIs[5] and a browsable web interface[6] for authors who wish to explore the object data by measure, magnitude range, familiarity, and/or measure variance.

AUTOMATED CONCRETE RE-EXPRESSION TOOLS

We have implemented adding familiar context, reunitization, and proportional analogy.strategies as automated concrete re-expression tools that use our database to help people reason about measurements. familiar context, (2) reunitization, and (3) proportional analogy. Given an unfamiliar measurement as input, each of our tools performs an energy minimization optimization to select the most effective re-expression object(s). Specifically, each tool defines an energy function $E(x)$ of synset x, that is a linear combination of terms related to the re-expression effectiveness criteria (see Section on Criteria for Effective Re-expression Objects).

Energy Terms and Effectiveness Criteria

Our re-expression tools use energy functions to select objects with high object-measure familiarity, low measure variance and high measure closeness. We describe energy terms for each criteria, where a smaller value for each term represents a better performing synset.

Object-Measure Familiarity. We define the object-measure familiarity energy term, E_{omf} of a synset x given an input measurement i as

$$E_{omf}(x) = \frac{crowdFam(i,x)}{20}$$

where $crowdFam(i,x)$ is the crowd familiarity score (i.e., the mean crowd rank) for synset x for the reference measurement from the crowd familiarity task with lowest absolute distance to i. We divide this score by 20 to normalize to a 0 to 1 range (recall that the penalty score for unranked objects is 20).

[5]https://github.com/jhullman/concrete-measurement-re-expressions
[6]http://measurements-interface.us-west-2.elasticbeanstalk.com/

Measure Variance. We define the normalized measure variance, $E_{mv}(x)$ of synset x as

$$E_{mv}(x) = \sigma_{x,m}/\mu_{x,m}$$

where m is the measure (weight, height, length, or volume) corresponding to i, and $\sigma_{x,m}$ and $\mu_{x,m}$ are the standard deviation and mean of synset x for measure m. Dividing the standard deviation by the mean (i.e., the coefficient of variation) allows for more meaningful comparison of the standard deviations for objects that have very different means [36]. Low values of E_{mv} occur when all of the measurements across the instance of x are similar to one another.

Measure Closeness. We define two energy terms computing the difference between the unfamiliar input measurement i and the corresponding measurement of synset x. The first term E_{rd} considers the rank distance between them as

$$E_{rd}(x) = |rank(i,x)|$$

where $rank(i,x)$ is the rank of x relative to the rank of i in sorted order for the corresponding measure. Suppose i is 55 inches and we consider synsets *park bench* (40 inches) and *writing desk* (42 inches). After sorting these two synsets by length, their rank distance with respect to the input measurement is *park bench* ($E_{rd} = 2$) and *writing desk* ($E_{rd} = 1$). This term prioritizes synsets with similar measurements to the input regardless of whether the measurement is smaller or larger. The penalty increases linearly with rank.

The second measure closeness term (for use in reunitization) considers the multiplier factor $mult(i,x) = i/x_m$, where x_m is the measurement for synset x corresponding to measure m. This factor is required to convert the synset measurement x_m to the input measurement i. We design the energy term to prioritize multipliers that fall within specific ranges that are known to be understandable to people based on number sense research [6, 17, 20]. Specifically, we define the multiplier term E_{mult} of synset x as

$$E_{mult}(x) = \begin{cases} 1/mult(i,x_m) & \text{if } 0 \le mult(i,x_m) < 1 \\ 0 & \text{if } 1 \le mult(i,x_m) < 3 \\ 0.1 & \text{if } 3 \le mult(i,x_m) < 10 \\ \frac{mult(i,x_m)-10}{2} + 1 & \text{if } 10 \le mult(i,x_m) \end{cases}$$

Multipliers between 1 and 3, which are easiest for people to understand [20], incur no penalty. Multipliers between 3 and 10 incur a small constant penalty as they require slightly more effort to understand, and multipliers greater than 10 also incur a linearly increasing penalty as they require mentally adding lots of copies of the re-expression object. Multipliers less than one incur a heavy non-linear penalty as they require imagining less than one whole object.

Object Similarity. We define the object similarity energy term $E_{sim}(x,y)$ for a pair of synsets x and y as

$$E_{sim}(x,y) = 1 - \frac{1}{path(x,y)} \quad (1)$$

where $path(x,y)$ is the length of the shortest path between x and y in the WordNet hyponym/hyponym graph [7, 31].

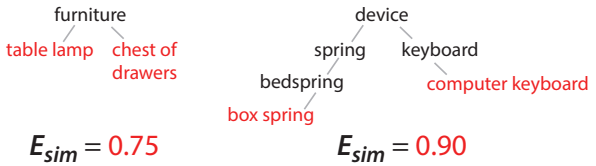

Figure 6. Object similarity describes how similar two objects are to one another based on the path length in the hypernym/hyponym graph of WordNet. Synsets *table lamp* and *chest of drawers* are connected by a shorter path than *computer keyboard* and *box spring*, and the corresponding value of the object similarity energy term is also lower.

Table 1. Adding familiar context top ranked objects for inputs that span a range of scales (0.1 to 1000) for three of the measures covered in our database.

	Weight (lb)	Length (ft)	Volume (gal)
Inputs	**Synset**	**Synset**	**Synset**
0.1	golf ball spring balance tea ball	straight pin safety valve cup hook	compass watch printer
1	football music box umbrella	mixing bowl tea chest dagger	soup bowl gravy boat football
10	starter chain saw vacuum	box spring vehicle tank	watering can bin electric heater
100	bench press china cabinet shed	tank submarine boat	chest of drawers writing desk box spring
1000	grand piano car vehicle	boat ship bridge	bench press shed aircraft

Synsets that are conceptually similar to one another are generally connected by shorter paths in this graph (Fig. 6). We take the reciprocal of the path length and subtract from 1 to ensure that shorter paths produce lower penalties.

Strategy: Adding Familiar Context

Our tool for adding familiar context is based on an energy function composed of three energy terms designed to evaluate the object-measure familiarity, measure variance and rank distance for each synset x as

$$E(x) = w_{omf}E_{omf}(x) + w_{mv}E_{mv}(x) + w_{rd}E_{rd}(x)$$

where the weights w control the strength of each term. To set the values of these weights, we use a typical process for manually setting weights in multicriteria optimization [11]. We initially set the weights based on the importance of each term, giving the highest weight to w_{omf}since familiarity is critical for the re-expression to help the person relate to the unfamiliar measurement, followed by w_{mv} then w_{rd}. We weight the rank distance the least as the penalty grows steeply as the synset gets further from the input measurement. We then used repeated experimentation to adjust the weights. We iteratively generated results for a wide range of input measurements, evaluating how many high ranking objects seemed effective as re-expression objects on each iteration. Using this procedure we set w_{omf} to 12, w_{mv} to 5, and w_{rd} to 1.

Given an unfamiliar input measurement i, we compute the energy for all of the synsets in our database. We select the n lowest energy synsets to serve as context objects, such that each new context object is at least a distance k from the objects previously added to the context set. This threshold prevents the familiar context objects from being clustered at a single value. We define k as a percentage of i. We use k=10% for all results that we present. Table 1 shows the top 3 (lowest energy) context objects for several types of input measures at different scales. The objects are generally familiar at each scale and differ depending on measurement type (weight, height, etc.) and scale. Results for a larger set of inputs, including the mean measurements of the re-expression objects, are presented in the supplemental material.

Strategy: Reunitization

Our automated reunitization tool defines the energy function for each synset x as

$$E(x) = w_{omf}E_{omf}(x) + w_{mv}E_{mv}(x) + w_{mult}E_{mult}(x).$$

Compared to the Adding Familiar Context function, this function substitutes the multiplier term, which aligns with the presentation of a reunitization, for the rank distance term. We find the specific term weights for w_{omf}, w_{mv}, and w_{mult} using the same procedure we used for our adding familiar context strategy. We set w_{omf} to 5, w_{mv} to 2, and w_{mult} to 4.

Given an unfamiliar input measurement i, we compute the energy for all synsets in our database and select the lowest energy synset as our reunitization object. Because people find it easier to understand numbers that contain fewer decimal places, with whole numbers being easiest [4, 9, 34] we round multipliers to the nearest integer. However, we also calculate the error introduced by the rounding and if the error is greater than 5% we retain as many additional decimal places as are needed to keep error below 5% in presenting the multiplier.

Table 2 shows the top 3 (lowest energy) reunitization objects for several types of input measurements and different scales. The results include synsets for many familiar household objects (i.e., a microwave, a book, pliers). Roughly two-thirds (36) of the 60 re-unitized measurements have a multiplier in our desired range of 1 to 3, a few (3) have a multiplier less than 1, and 0 have a multiplier greater than 10. Results for more inputs, including the mean measurements for the re-expression objects, are presented in the supplemental material.

Strategy: Proportional Analogy

Our proportional analogy tool defines an energy function for a pair of synsets x and y:

$$E(x,y) = w_{omf}E_{omf}(x,y) + w_{mv}E_{mv}(x,y) + w_{sim}E_{sim}(x,y)$$

where the pairwise familiarity and measure variance terms are:

$$\begin{aligned} E_{omf}(x,y) &= E_{omf}(x) + E_{omf}(y) \\ E_{mv}(x,y) &= E_{mv}(x) + E_{mv}(y) \end{aligned}$$

The object similarity term $E_{sim}(x,y)$ computes similarity between the synsets. We set $w_{omf} = 2$, $w_{mv} = 5$, and $w_{sim} = 20$.

Given a pair of unfamiliar input measurements i and j, it may not be possible to identify a pair of synsets in our database

Table 2. Reunitization top ranked objects with multiplicative factors for inputs that span various scales (0.1 to 1000) for weight, height, length, and volume.

	Weight (lb)	Length (ft)	Volume (ft)
Inputs	**Expression**	**Expression**	**Expression**
	1.1 golf ball	0.6 straight pin	3.3 shot glass
0.1	2.5 cup hook	0.6 guitar pick	3.3 golf ball
	1.4 stylus	0.5 cup hook	1.7magnet
	1.3 soccer ball	1.1 hammer	2.3 sugar bowl
1	2.6 water bottle	1.3 bicycle seat	1.4 travel iron
	1.4 headset	1.6 icepick	2.9 butter dish
	2.3 laptop	1.5 box spring	1.6 record player
10	1.3 power drill	1.7 double bed	1.5 aquarium
	1.3 cash register	4 sword	6.2 sieve
	1.4 secretary desk	7 car	9.8 bin
100	1.1 bench press	4.7 tank	4.8 toy box
	1.1 platform bed	6.8 vehicle	6.3 microwave
	4.7 gas range	7.5 submarine	2.4 hot tub
1000	6.6 shed	2.3 ship	1.5 shed
	5.4 hot tub	4.2 boat	5.5 wardrobe

that have the same measurement ratio as i to j. Therefore, we identify all pairs of synsets whose measurement ratio is within 5% of the input ratio, and compute the energy for each such synset pair (x, y). We choose the lowest energy pair for the analogy. Table 3 shows the top 3 (lowest energy) analogy pairs of objects and their error relative to the input ratio for a set of ratios for the measure weight.The pairs of objects in the analogies are often related through a common hypernym due to our object similarity term $E_{sim}(x, y)$ (e.g., console table and breakfast table are both tables).We provide results for a larger set of inputs are presented in the supplemental material.

Does learning weights from humans improve results?

An alternative method for setting the term weights in our automated re-expressions is to have people rank a large set of re-expressions and then use their preferences to learn the weights. To evaluate how human-learned weights improves the performance of our reunitization tool, we compared our reunitization algorithms ranking of re-expressions for each measure to a ranking of the same results that we generated using term weights optimized from a human-ranked gold standard set. Overall, we found that while the optimal human-learned weights improve the results slightly, the gains for each measure are relatively small. We also use the human ranked results to evaluate the impact of the different terms in our Reunitization energy function, and find that each term contributes positively. We detail this analysis in supplemental material.

APPLICATIONS

We have developed separate Web-based applications for each of our automated concrete re-expression tools that apply the tools to measurements in text articles. We devised several guidelines for presenting re-expressions in reading contexts and used these to inform our design: 1) the application should make clear when a re-expression is available for a measurement, 2) users should be able to request re-expressions on demand, 3) reunitization multipliers should balance preciseness with readability, and 4) re-expressions should be expressed using visuals and text. Users click on the measurements to view re-expressions, which appear in a sidebar (Figs. 1, 2, 7).

Table 3. Proportional analogy top ranked objects for input ratios that span a range of scales (1:1.5 to 1:1000) for weight and volume. Error indicates the percent error between the input ratio of measurements and the ratio between the measurements of the analogy objects.

	Weight	
Inputs	**Synset**	**Error**
	console table : breakfast table	0.009
1:1.5	skateboard : handcart	0.033
	microwave : box spring	0.015
	console table : chest of drawers	0.0034
1:2	double bed : platform bed	0.022
	cd drive : stringed instrument	0.013
	skateboard : hand truck	0.0056
1:4	charger : router	0.011
	file server : personal computer	0.001
	laptop : box spring	0.0024
1:10	router : fan	0.0032
	fire alarm : electric fan	0.0035
	cd drive : chest of drawers	0.00047
1:100	hammer : bench press	0.00027
	hammer : platform bed	0.0004
	knife : gas range	0.0000
1:1000	cup hook : console table	0.0000
	stylus : secretary	0.0000

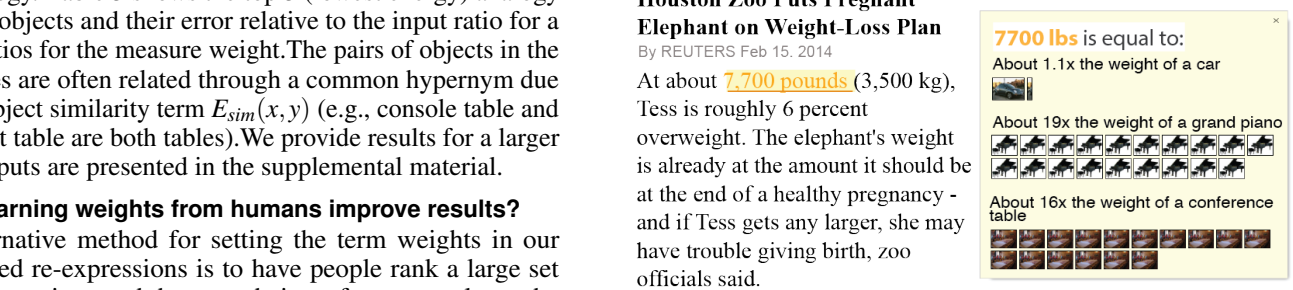

Figure 7. An example of our reunitization application. When the user selects a measurement in the text, the concrete re-expression is presented in the right sidebar.

Tagging Measurements in Text Articles: Given the url of a text article as input, our applications tag all weights, heights and lengths, and volumes using a custom regular expression parser to match variations on standard terms for units (e.g., [mg, g, kg, lb, ton] for weight, [mm, cm, in, ft, m, km, mi], for height or length, and [mm^3, cm^3, in^3, l, gal, ft^3, m^3, acre-foot] for volume.We render tagged measurements in an orange font.

Adding Familiar Context: Clicking on a tagged measurement opens a sidebar with a vertical number line that depicts the n lowest energy synsets (n is set to 5 in all figures) as returned by our automated re-expression tool (Figs. 1 left, 7 top).

Reunitization: Clicking on a measurement populates a sidebar with a horizontal pictograph bar chart composed of multiples of representative ImageNet synsets for each of the five lowest energy synsets returns by our reunitization tool (Fig. 1 middle,7 center). The "bars" are easily comparable based on a constant image width and a fixed left axis from which all bars extend. The text below each bar describes the multiplier for converting between the familiar object measurement and the measurement from the article. People find it easier to un-

derstand numbers that contain fewer decimal places; whole numbers are easiest [4, 9, 34]. Therefore our policy is to round these multipliers to the nearest integer. However, we also calculate the error introduced by the rounding and if the error is greater than 5% we retain additional decimal places in presenting the multiplier. This policy adaptively maintains more precision when the original measurements are small and ensures that the error is never higher than 5%.

Proportional Analogy: Clicking on a pair of measurements of the same type (e.g. both weights) opens a sidebar showing a text representation of the analogy (Fig. 1, 7 bottom).

EVALUATION

To evaluate our automated re-expression tools, we first consider how their coverage, familiarity, and multiplicative factors compare to alternative solutions for generating reunitizations. We then consider the value of the re-expressions for measurement comprehension through controlled online studies with crowdworkers. In a first study, we consider whether viewing re-expressions for measurement helps users understand new measurements. In a second set of studies, we consider whether users perceive the re-expressions as helpful for understanding measurements in news articles.

Comparison to Existing Automated Reunitization Tools

The Dictionary of Numbers (DN), The Measure of Things (MT), and Wolfram's Alpha (WA) can present reunitizations as text for some measures. We compare our reunitization tool against these tools for a set of 60 input measurements (15 measurements for each of the four measures, spanning 8 orders of magnitude and the halfway points between them from 0.001 to 10,000). We considered coverage (e.g., how many input measurements a tool was able to generate re-expressions), the familiarity of the objects used in re-expressions, and the magnitude of the multipliers used in the re-expressions. Overall, we find that our tool has better coverage than DN and WA, uses more unique objects across the set of results than DN or WA, is much more likely to return re-expressions using everyday objects than DN, WA, and MT, and is much more likely to use multipliers between 1 and 10 than DN, WA, and MT. We describe the analysis and results in supplemental material.

User Studies of Comprehension & Perceived Helpfulness

Do our tools help people understand measurements better?

We conducted a within-subjects user study on Amazon Mechanical Turk to answer the question, *Does viewing our concrete re-expressions help users estimate new measurements more accurately?* Participants were shown images of objects and asked to guess their weight, height, length, or volume, with and without the aid of our re-expressions. We summarize the study below and in detail in supplemental material.

Stimuli and Procedure: We adapted the estimation study used by Barrio et al. [3], in which participants must provide a missing measurement and get either no re-expressions or dynamically generated re-expressions as they enter their guess. We selected 8 synset instances (Amazon products) from our database (e.g., a chest of drawers, gas range, grand piano, etc.). We created 16 trials, to be completed by each subject in a single HIT. Each trial presented an image of the object and synset name, and asking participants to provide the weight, height, length, or volume by positioning a slider handle, either with or without the benefit of viewing re-expressions as feedback for the guesses they make. *No Re-expression* trials presented the slider with no dynamic re-expressions. For Re-expression trials, we varied whether the re-expressions used the *Adding Familiar Context* or *Reunitization* strategy between subjects. We randomly chose a slider range between 1.25 and 5 times the true measurement for each trial to vary the answer position.The 16 trial HIT carried a reward of $2.00 and bonus of $1.00 ($\mu$=$7.20 per hour).

Results: 120 workers completed the task. We use mean absolute percentage error, i.e., $\frac{|response-true|}{true}$ as our error measure. Error was lowest for Adding Familiar Context (μ: 0.285 σ: 0.243), followed by Reunitization (μ : 0.347 σ: 0.282), then "No Re-expression" (μ : 0.35 σ: 0.32). To account the repeated measures design, we ran a mixed effects linear regression to regress the absolute error term on treatment (Adding Familiar Context, Reunitization, No Re-expression). We find that viewing an Adding Familiar Context re-expression reduced error by 0.08 (i.e., 8%; 95% CI: [5, 12]) relative to No Re-expression. Viewing a Reunitization re-expression did not reliably reduce error (95% CI contains 0: [-3, 3]). The cause may be the effort to interpret the re-expression objects as a multiple of the slider input, which changes with every slider move.

Do users find re-expressions useful when reading news?

We conducted two repeated measures between-subjects user studies on Amazon Mechanical Turk to answer the question, *Do users find our concrete re-expressions helpful for understanding measurements in text?* A first study compared people's ratings of the helpfulness of news article content for understanding measurements across our three strategy implementations and No Re-expression conditions. A second follow-up study controlled for the possibility that subjects' higher ratings were driven by their observation of a difference between No Re-expressions and our strategy implementations, rather than re-expression quality. We compared our Adding Familiar Context and Reunitization strategies to *Random* re-expressions: re-expressions using objects that were randomly selected from our database of objects.

Stimuli and Procedure: To compare our strategy implementations to no re-expressions, we collected 6 text articles containing multiple measurements such as weights, heights, lengths, and volumes from news outlets. We selected 10 single measurements (e.g., "1000 ft") and 5 pair of measurements (e.g., weights of 200 lbs and 1000 lbs) for proportional analogies from the articles. These 15 measurements became separate trials in a single Mechanical Turk HIT. For each trial a worker was randomly assigned to either *No Re-expression*, *Adding Familiar Context* or *Reunitization* (for single measurements) treatment, or *Proportional Analogy* (for paired measurements). We visually highlighted the measurement(s) for each trial. We asked workers to examine the article content and decide *Is the article content helpful for understanding the highlighted measurement? (Yes/No)*. Workers were then asked to rate their agreement with the sentence: *The content of the article helped me to understand the size of the highlighted measurement.*, using a 7 point Likert scale from Strongly Disagree to

Strongly Agree. We also provided a text box and asked them to *Briefly describe how the article content is or is not helpful for understanding the size of the measurements.* To ensure that they paid attention, workers also had to specify the highlighted measurement for each trial. The HIT carried a reward of $1.50 ($\mu$=$15.33 per hour).

Results: Of 465 total trials, we analyzed 453 (from 31 total workers) where the worker had correctly identified the highlighted measurement (No Re-expression: 236,Reunitization: 76, Adding Familiar Context: 71, Proportional Analogy: 70). The mean helpfulness Likert rating was highest for Adding Familiar Context (μ: 5.1 σ:1.9), followed by Proportional Analogy (μ : 4.7 σ: 2.2), followed by Reunitization (μ:4.5 σ: 2.1), then No Re-expression (μ:3.6 σ: 2.1).

To account the repeated measures design, we ran mixed effects logistic regressions to compare the probability that a worker would say "Yes" when asked if the article content was helpful across conditions. We ran mixed effects linear regressions to assess differences in helpfulness ratings across conditions. For each analysis, we ran one model for the single measurement conditions (Adding Familiar Context, Reunitization, No Re-expression) and one for the paired measurement conditions (Proportional Analogy, No Re-expressions). All models regressed the responses to Question 1 or Question 2 on treatment and an indicator of trial order, and included worker id as a random effect. Full analyses are in supplemental material.

Compared to seeing No Re-expression for a single measurement, a worker who saw an Adding Familiar Context re-expression was 20.6 times more likely to answer "Yes" to Question 1 (logit scaled 95% CI on effect estimate: [2.0, 4.0]), and on average rated the helpfulness of the content 2.1 pts (out of 7) higher (95% CI: [1.6, 2.6]). Compared to No Re-expression, a worker who saw a Reunitization re-expression was 18 times more likely to answer "Yes" to Question 1 (logit scaled 95% CI on effect estimate: [2.0, 3.8]), and on average rated the helpfulness of the content 1.9 pts (out of 7) higher (95% CI: [1.4, 2.4]). We saw no reliable difference in the odds ratios or helpfulness ratings between Adding Familiar Context and Reunitization (95% CIs on effect estimates contained 0).

Compared to seeing No Re-expression for paired measurements, a worker who saw a Proportional Analogy re-expression was 9.3 times more likely to answer "Yes" to Question 1 by (logit scaled 95% CI on effect estimate: [1.3, 3.1]), and on average rated the helpfulness of the content 2.1 pts (out of 7) higher (95% CI: [1.5, 2.7]).

Reasons that workers provided for helpful re-expressions for Question 2 were familiarity of the object, familiarity with the object's measure, and seeing multiple re-expressions.

Follow-Up Comparison with Random Re-expressions: To further evaluate the impact of using our energy functions to rank re-expressions generated from objects in our database, we reran our usefulness study for single measurements but where "Random" re-expressions–re-expressions created by randomly drawing five objects to serve as context objects or reunitization objects, respectively–replaced the No Re-expression trials. Seeing either an Adding Familiar Context or a Reunitization re-expression increased the probability that a worker would answer "Yes" to Question 1 by 14.3 times (95% CI: [6.3, 33.3]) and increased the worker's helpfulness rating by an average of 1.9 pts (out of 7) (estimated 95% CIs: [1.5, 2.4]). We report full methods and results in supplemental material.

DISCUSSION & LIMITATIONS

We rely on Amazon, DBPedia and Freebase to build our object database. Each source has biases in the objects it includes. This results in biases in the measure distributions we obtain (e.g., at some magnitudes the database is dense with many synsets while at others, such as very small measurements, it is sparse).

Our automated strategy implementations select objects using a familiarity model that aggregates preferences obtained from U.S. based users. However, individual familiarity with objects will vary. Modelling individual and cultural factors (e.g., using individual online purchase histories, or geolocation information as a signal of culture) may improve the usefulness of re-expressions. Applications that enable customization of term weights, or that learn preferred weights over time, may also help address individual preferences among users.

Another form of personalization is to tailor the re-expressions to the context in which they appear, such as the topic of the text in an article. In contrast to existing automated solutions that rely on a small collection of objects and reuse them frequently, the larger database that we create could support customizing re-expressions based on context.

We rely on WordNet to represent objects for re-expressions and focus on hyponyms of the synset *artifact* in building our database, because properties of these synsets aligned with object criteria like being concrete and countable. However, this restricts our approach from considering other objects that might be useful in re-expressions, like natural objects. For example, a *snowflake* or a *giant redwood* are natural objects that could be used to produce re-expressions for very small or large measurements that are familiar to many users.

We chose to use the mean measurement across object (synset) instances after observing through self-experimentation that mean, median, and geometric mean performed similarly. As a result, our database supports the possibility of reporting uncertainty with the expected value of a measurement. For example, rather than presenting a rounded multiplier in our Reunitization application, we might present a range.

CONCLUSION

We presented a set of tools for automatically re-expressing unfamiliar measurements using the measurements of familiar objects. The key idea of our approach is to identify criteria for effective re-expressions then build a database of familiar objects and their measurements by combining information from semantic databases, object databases and crowdsourcing. We show the database can be used to implement three common re-expression strategies. Our tools make it easier for publishers, educators, or journalists to enhance their audience's understanding of measurements through concrete measurement re-expressions in various informal and educational contexts.

ACKNOWLEDGEMENTS

This work was partially supported by a Tableau Fellowship.

REFERENCES

1. Amazon. 2014. Amazon Product Affiliate API. https://affiliate-program.amazon.com/gp/advertising/api/detail/main.html. (2014).
2. Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2016. Are Elephants Bigger than Butterflies? Reasoning about Sizes of Objects. *arXiv preprint arXiv:1602.00753* (2016).
3. Pablo J Barrio, Daniel G Goldstein, and Jake M Hofman. 2016. Improving comprehension of numbers in the news. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2729–2739.
4. Susana Bautista, Raquel HervÃąs, Pablo GervÃąs, Richard Power, and Ra Williams. 2011. How to Make Numerical Information Accessible: Experimental Identification of Simplification Strategies. (2011).
5. Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proc. of SIGMOD '08*. ACM, New York, NY, USA, 1247–1250.
6. Elizabeth M Brannon. 2006. The representation of numerical magnitude. *Current opinion in neurobiology* 16, 2 (2006), 222–229.
7. Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Comput. Linguist.* 32, 1 (March 2006), 13–47.
8. Lou Burnard. 1995. Users Reference Guide British National Corpus Version 1.0. (1995). http://www.natcorp.ox.ac.uk/
9. Jamie I. D. Campbell. 2005. *Handbook of Mathematical Cognition*. Psychology Press.
10. Arun Tejasvi Chaganty and Percy Liang. 2016. How Much is 131 Million Dollars? Putting Numbers in Perspective with Compositional Descriptions. *arXiv preprint arXiv:1609.00070* (2016).
11. Vira Chankong and Yacov Y Haimes. 2015. Multiobjective Decision Making: Theory and Methodology. (Sept. 2015). http://store.doverpublications.com/0486462897.html
12. Fanny Chevalier, Romain Vuillemot, and Guia Gali. 2013a. Using Concrete Scales: A Practical Framework for Effective Visual Depiction of Complex Measures. *IEEE TVCG* 19, 12 (2013), 2426–2435.
13. Fanny Chevalier, Romain Vuillemot, and Guia Gali. 2013b. Using Concrete Scales: A Practical Framework for Effective Visual Depiction of Complex Measures (Presentation). (2013). Presented at IEEE InfoVis 2013.
14. Glen Chiachieri. 2014. The Dictionary of Numbers. http://www.dictionaryofnumbers.com/. (2014).
15. Jonathon Clase. 2017. The Measure of Things. http://www.bluebulbprojects.com/measureofthings/default.php. (2017).
16. DBpedia. 2015. DBpedia: Downloads 2015-10. (2015). http://wiki.dbpedia.org/Downloads2015-10
17. Stanislas Dehaene, Elizabeth Spelke, and Lisa Feigenson. 2004. Core Systems of Number. *Trends in Cognitive Sciences* 8, 7 (2004), 307–314.
18. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.
19. Ari Rappaport Dmitry Davidov. 2010. Extraction and Approximation of Numerical Attributes from the Web. (2010).
20. Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. 2004. Core systems of number. *Trends in cognitive sciences* 8, 7 (2004), 307–314.
21. Google. 2014. Freebase API (Deprecated 2016): Data Dumps. (2014). https://developers.google.com/freebase/
22. Mark F Horstemeyer. 2009. Multiscale modeling: a review. In *Practical aspects of computational chemistry*. Springer, 87–135.
23. M. Gail Jones, Grant E. Gardner, Amy R. Taylor, Jennifer H. Forrester, and Thomas Andre. 2012. Students' Accuracy of Measurement Estimation: Context, Units, and Logical Thinking. *School Science and Mathematics* 112, 3 (March 2012), 171–178.
24. M. Gail Jones, Manuela Paechter, Chiung-Fen Yen, Grant Gardner, Amy Taylor, and Thomas Tretter. 2013. TeachersâĂŹ Concepts of Spatial Scale: An international comparison. *International Journal of Science Education* 35, 14 (Sept. 2013), 2462–2482.
25. M. Gail Jones and Amy R. Taylor. 2009. Developing a sense of scale: Looking backward. *Journal of Research in Science Teaching* 46, 4 (2009), 460–475.
26. Yea-Seul Kim, Jessica Hullman, and Maneesh Agrawala. 2016. Generating personalized spatial analogies for distances and areas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 38–48.
27. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2014. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* (2014).
28. Alejandra Magana, Sean Brophy, and Lynn Bryan. 2012. An Integrated Knowledge Framework to Characterize and Scaffold Size and Scale Cognition (FS2C). *Int'l J. of Science Educ.* 34, 14 (2012), 2181–2203.
29. Alejandra J Magana. 2014. Learning strategies and multimedia techniques for scaffolding size and scale cognition. *Computers & Education* 72 (2014), 367–377.

30. George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (Nov. 1995), 39–41.

31. NLTK. 2015. NLTK Python WordNet Interface. (Sept. 2015). http://www.nltk.org/howto/wordnet.html

32. John Allen Paulos. 1988. *Innumeracy: Mathematical Illiteracy and Its Consequences*. Macmillan.

33. Chris Riederer, Jake M. Hofman, and Daniel G. Goldstein. 2018. To put that in perspective: Generating analogies that make numbers easier to understand. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA.

34. Bruce M. Ross and Trygg Engen. 1959. Effects of round number preferences in a guessing task. *Journal of Experimental Psychology* 58, 6 (1959), 462–468.

35. Charles Seife. 2010. *Proofiness: The Dark Arts of Mathematical Deception* (1 ed.). Viking Adult.

36. Barbara G. Tabachnick and Linda S. Fidell. 2006. *Using Multivariate Statistics (5th Edition)*. Allyn & Bacon, Inc., Needham Heights, MA, USA.

37. Thomas R Tretter, M Gail Jones, Thomas Andre, Atsuko Negishi, and James Minogue. 2006. Conceptual boundaries and distances: Students' and experts' concepts of the scale of scientific phenomena. *Journal of research in science teaching* 43, 3 (2006), 282–319.

38. Michael J Wilber, Iljung S Kwak, and Serge J Belongie. 2014. Cost-effective hits for relative similarity comparisons. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

39. Wolfram|Alpha. 2015. WolframAlpha Computational Knowledge Engine. (2015). http://www.wolframalpha.com/ Last visited on 08/4/2015.