On the Computation of Semantically Ordered Truth Values of Linguistic Protoform Summaries

Akshay Jain*, *Student Member, IEEE,* and James M. Keller†, *Life Fellow, IEEE*
Electrical and Computer Engineering
University of Missouri
Columbia, MO, USA
*aj4g2@mail.missouri.edu, †kellerj@missouri.edu

***Abstract*—Linguistic summaries provide a promise for extracting useful information from large datasets and present them in the form of natural language. Their applications can be found in various disciplines such as eldercare, financial time series etc. In this work we focus on linguistic protoform summaries and point out at some problems associated with the truth value computation methods present in literature. We develop a technique which produces more intuitive truth values for three different kinds of linguistic protoform summaries and illustrate this with help of some examples. Moreover, we show through a mathematical proof that our method is very robust and the truth values always follow the semantic order of the language they are representing.**

Keywords—linguistic summaries, truth value, fuzzy logic, data to text generation.

I. INTRODUCTION

As the amount of data increases, converting it into useful information becomes an increasing difficult problem. This has been widely discussed in the data science community under the name of big data which is said to be comprised of the four V's – Volume, Variety, Velocity and Veracity [3]. Each of these V's pose a unique set of problems in terms of understanding the information content. For instance, a widely used approach in order to get to know the data well is to visualize it. However, as the variety/dimensionality of the data increases visualization becomes more difficult. Another technique is to compute some statistical metrics to get a big picture of the nature of the data. Though the statistical metrics like mean, median etc. provides some insight into the information content, they might neglect a lot of intrinsic details that may be very important. Moreover, it may be difficult for an untrained eye to analyze these metrics to keep up with the data. Linguistc summaries offer a human-centric method to compress and explain large amounts of raw data. For example, [4] discuss the implications of vast amounts of data generated in Neonatal Intensive Care Units and how textual summaries can help the doctors, nurses and caretakers of infants make better use of it. Linguistic summaries hold a promise of providing a solution to these problems by describing the information content in human readable natural language. Along with the capability of describing intricate information in heterogeneous records, it naturally makes it easy for the people with different background to analyze it, for example in case of monitoring sensor output from the home of elderly residents [5].

A number of ways to generate textual summaries of numeric data have been presented in the past. Two prominent approaches are the statistical rule based methods [4, 6] and fuzzy logic techniques [1, 2, 5, 7-13]. In [14] the authors point that statistical rule based techniques produce more sophisticated and longer text summaries of data, while are not so flexible in data processing due to the hard coding of the data selection rules as compared to the fuzzy logic techniques

Largely, three types of Linguistic Protoform Summaries (LPS) have been talked about in the literature, with each focusing on a different property of the data. They can be exemplified by *Many balls are big*, *Some balls are light and small* and *Few of the red balls are small*. We call these simple protoform, protoform 2 and protoform 3 summaries, respectively. For each summary one or more metrics are computed indicating its validity with respect to the underlying data. In this work we focus on computation of the metric associated with these three types of summaries.

The summaries of the form *Many balls are big* were introduced by Yager [1] as a way to summarize both numeric and non-numeric data. Yager's LPS (and our work) basically is comprised of three terms: a summarizer (e.g., *big, small, red* etc), a quantifier (e.g., *Few, Some, Many* etc) and a truth value. Since then this idea has been taken forward both in terms of new protoforms and the way the truth value of a summary is computed. The quantifier forms an integral part of the summaries describing the presence of one or more attributes of data. Historically, a lot of focus has been put on evaluating the sentences based on the type of quantifier used. Generally, they can be used to describe precise information (such as *all, none* etc.) or can be of imprecise nature (such as *Few, Some* etc.). The use of fuzzy sets in order to model quantifiers conveying non-crisp information seems to be adequate. Fuzzy quantification is a widely studied topic with many approaches to evaluate the summaries involving fuzzy quantifiers. Authors in [15] provide an in depth analysis of the different quantification techniques. This work builds on a methodology using the Sugeno integral [16]. In the following we present an overview of the development of different aspects of LPS.

In [10] the authors modified the structure of protoforms to characterize different aspects of time series data, namely, how long certain trends remain constant, and how much and how fast they change. They fused the information using the Sugeno Integral [16] to compute truth values of these protoforms. Then in [9], they introduced a metric called degree of focus to reject summaries which do not add much information. An approach to compare two time series was proposed in [8, 17]. They generated

LPS to depict the changes in two time series at various levels. In [5] motion and restlessness data for an elderly participant is analyzed using LPS. Then in [18] a distance metric for a space of linguistic summaries was developed. This allowed the authors to compute distances between groups of linguistic summaries in the eldercare sensor data. The summaries could then be clustered and the clusters eventually represented by Linguistic Medoid Prototypes [13]. Conceptually, the goal of that work was to utilize LPS as a compact but understandable feature set for heterogeneous data streams, both for the purpose of communication and automatic detection of anomalies [19]. A different take to produce summaries at various granularities was proposed in [20] with a method similar to Yager [1] to compute truth values. Reference [21] focusses on summaries of the form *Y's are P and R*, e.g., *Employees are well paid and young*. They call the metric associated with the summaries as the degree of truth, which indicates the percentage of objects satisfying all of the given attributes. The method to compute the degree of truth is somewhat similar to [10], however their main concern is agreement of different attributes rather than the use of quantifiers. In [22, 23], genetic algorithms are used to search for interesting summaries of data. Summaries with features such as high truth value, compactness and uniqueness are deemed as interesting.

The authors of [2] have shown that the Zadeh calculus for defining truth values, (defined and analyzed in Section II.B), can produce very non-intuitive values [2]. In that paper, a method to address the main problem was introduced, but issues with respect to a natural interpretation of truth values still remained. We build on it to produce truth values which are more intuitive with respect to the semantic meaning of the quantifiers. The truth value of an LPS represents the amount of information it contains or how well it represents the underlying data. In Section II we discuss few of the techniques to compute truth values that are relevant to our work. Section III presents our approach with focus on its benefits accompanied by a theorem proving the robustness of our method. In Section IV, we extend the proposed method for summaries with extended protoforms, e.g., *Few of the balls are red and small*, *Few of the red balls are small*.

II. BACKGROUND

LPS are short natural language sentences whose form is prescribed depending on the application and the nature of data to be summarized. A general simple form of linguistic protoform summary is

A y's are P

for example, *Few balls are big.* Here, *A* is the linguistic quantifier which is selected according to the quantity of the objects being summarized, like *most, many, few, some* etc. The variable *y* represents the type of the objects, like *balls, days etc. P* is the summarizer, encapsulating a feature, like *big, small, tall* etc. The truth value *T* is calculated for every summary representing its informativeness. There are several methods to compute truth values of which we describe two that are relevant to this work. The linguistic variables *A* and *P* are fuzzy sets over suitable domains and are represented by membership functions defined as normal, convex L-R fuzzy numbers [24] as shown in equation (1).

$$A(x)=\begin{cases} L\left(\dfrac{x-a}{b-a}\right) & if\ x\in[a,b] \\ 1 & if\ x\in(b,c) \\ R\left(\dfrac{d-x}{d-c}\right) & if\ x\in[c,d] \\ 0 & otherwise \end{cases} \quad (1)$$

Where, $L(x)$ and $R(x)$ are monotonically non-decreasing shape functions with; $L(0)=R(0)=0, L(1)=R(1)=1$. While L-R fuzzy numbers are general, many computational problems use either trapezoid functions or triangle functions (b = c) where *L* and *R* are linear. It may be the case that the value of the function is 1 at the left or right endpoint of the domain (for e.g., *Almost None* and *Many,* respectively in Figure 1). In such cases, by definition, $L\left(\dfrac{x-a}{b-a}\right)=1, if\ x=a=b$ and $R\left(\dfrac{d-x}{d-c}\right)=1, if\ x=c=d$. Also note that for the case of quantifiers, we only concentrate on the ones defined within the domain [0, 1]. However, there is no such restriction for summarizers.

A. Semantically Ordered Quantifiers

Words in languages have semantics associated with them. For instance, given a bag containing a number of balls of different sizes, the sentence *Few balls are big* corresponds to fewer number of big balls in the bag as compared to the sentence *Some balls are big*. In line with this, the membership functions of the linguistic quantifiers (*Few and Some*, respectively) should also follow the same semantic order. We define semantically ordered quantifiers to follow this reasoning. To fix ideas, suppose that $A(x)$ and $B(x)$ are the membership functions of the quantifiers like *Few* and *Some* in the above two sentences (as defined in equation (2) and (3) respectively).

$$A(x)=\begin{cases} L1\left(\dfrac{x-a}{b-a}\right) & if\ x\in[a,b] \\ 1 & if\ x\in(b,c) \\ R1\left(\dfrac{d-x}{d-c}\right) & if\ x\in[c,d] \\ 0 & otherwise \end{cases} \quad (2)$$

$$B(x)=\begin{cases} L2\left(\dfrac{x-p}{q-p}\right) & if\ x\in[p,q] \\ 1 & if\ x\in[q,r] \\ R2\left(\dfrac{s-x}{s-r}\right) & if\ x\in[r,s] \\ 0 & otherwise \end{cases} \quad (3)$$

Then we call these membership functions semantically ordered if,

$A(x)\geq B(x)\ \forall x\in[a,b]$, and $B(x)\geq A(x)\ \forall x\in[r,s]$

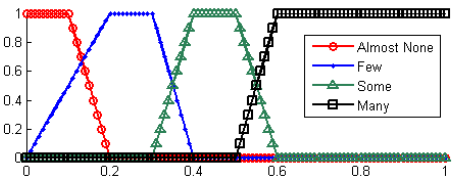

Figure 1: Membership functions of Quantifiers

Primarily, we avoid the cases where the functions overlap in an ambiguous fashion. The membership functions of the quantifiers shown in Figure 1 are semantically ordered in the following order: *Almost None, Few, Some, Many*. Trapezoidal functions are usually denoted as Trap(a,b,c,d).

B. *Truth value computation using Zadeh's calculus of quantified propositions [25]*

The first and still popular method used to compute truth value of LPS was introduced by Yager in [1] which employs Zadeh's calculus of linguistically quantified propositions [25]. It is given by equation (4).

$$T(A\ y's\ are\ P) = A\left(\frac{1}{N}\sum_{i=1}^{N} P(y_i)\right) \qquad (4)$$

where $A(x)$ and $P(x)$ are the membership functions of the quantifier and summarizer respectively, y_i is the i^{th} object in the data to be summarized, with N total number of objects. When the data is relatively crisp, this method produces intuitively correct truth values. However, due to the use of averaging operator, a discrepancy arises in case of fuzzier datasets. We illustrate this with the help of some examples.

Example 1

In this example we consider a scenario in which we would like to produce linguistic summaries of bags containing 10 balls of various sizes. We also assume that a separate process provides the summarizer value (bigness) of each ball varying from 0 to 1. Since the validity of LPS can be very subjective, the data in Table 1 is tailored to highlight the problems mentioned above.

Table 1: Bigness of each ball

	Bigness									
Bag 1	0.9	0.9	0.9	0.7	0.7	0.7	0.7	0.1	0.1	0.1
Bag 2	0.4	0.3	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.2
Bag 3	0.9	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3

We define four semantically ordered quantifiers as shown in Figure 1 and compute truth values of the four sentences – *Almost None of the balls are big, Few balls are big*, *Some balls are big* and *Many balls are big*. Note that the membership functions of these quantifiers are of the form presented in equation (1).

According to intuition, for bag 1, *Many balls are big* seems to be the best choice, while for bag 2 and 3, it is safe to say that out of the four summaries, *Almost None of the balls are big* best represents the data, and it should have the highest truth value. Because of the fuzziness of the data, the summaries which do not characterize the data best may have a truth value greater than 0. However, their truth values are expected to follow the semantic order of the quantifiers.

Table 2 displays the truth values produced using equation (4) for the data shown in Table 1. The truth value computed for summaries corresponding to bag 1 seems to be in line with the data. However, this is not the case for bags 2 and 3. First, even though in bag 2 none of the balls are really big, the truth value of *Few balls are big* is calculated as maximum. Second, in bag 3, the truth value of *Some balls are big* is computed to be maximum which does not correspond to the fact that only one ball is really big. As noted in [2], this discrepancy is due to the averaging of memberships in the formula, which works better with monotonically non-decreasing quantifiers (such as *Many*) as compared to others.

Table 2: Truth values using [1], of LPS associated with example 1

	Almost None	**Few**	**Some**	**Many**
Bag 1	0.00	0.00	0.20	0.80
Bag 2	0.00	1.00	0.00	0.00
Bag 3	0.00	0.40	0.60	0.00

C. *Truth value computation using method presented in [2]*

The problem associated with averaging of the memberships highlighted above was addressed recently by [2]. The reasoning behind their technique is that while summarizing a set of objects humans reject the objects with low memberships. To this end, they use the discrete form of Sugeno Integral [16] to compute the truth value of LPS of the form *A y's are P*, as shown in equation (5).

$$T(A\ y's\ are\ P) = \max_{\alpha}\left(\alpha \wedge A(P_\alpha)\right) \qquad (5)$$

where, ^ is the minimum operator, $P_\alpha = \frac{\left|\{y_i \in Y \mid P(y_i) \geq \alpha\}\right|}{N}$ is the proportion of objects whose membership in $P(x)$ is greater than or equal to α, $|.|$ denotes the cardinality of a set and $A(x)$ is a normal, convex and monotonically non-decreasing membership function of the quantifier A. Specifically, equation (5) is the Sugeno fuzzy integral with fuzzy measure defined by $A(x)$. This formulation satisfies the definition of a fuzzy measure as long as $A(x)$ is a monotonically non-decreasing function. Hence, as shown, it is not suitable to determine the truth values, say, for propositions *Almost None*, *Few* and *Some* in Figure 1. In the following we explain the procedure to calculate the truth value when the quantifiers are not monotonically non-decreasing.

Suppose we have quantifier A, with membership function $A(x)$ of the form shown in equation (1). In order to compute the truth value of the summary *A y's are P*, $A(x)$ is first split into $A_1(x)$ and $A_2(x)$. The function $A_1(x)$ is monotonically non-decreasing while $A_2(x)$ is a monotonically non-increasing function. In order to use equation (5) to compute truth value of a summary involving A_2, its complement, $\overline{A_2}(x)$ is defined as shown in equation (7).

$$A(x) \Rightarrow \begin{cases} A_1(x) = \begin{cases} 0 & \text{if } x < a \\ L\left(\frac{x-a}{b-a}\right) & \text{if } x \in [a,b] \\ 1 & \text{if } x > b \end{cases} \\ A_2(x) = \begin{cases} 1 & \text{if } x < c \\ R\left(\frac{d-x}{d-c}\right) & \text{if } x \in [c,d] \\ 0 & \text{if } x > d \end{cases} \end{cases} \quad (6)$$

$$\overline{A_2}(x) = \begin{cases} 0 & \text{if } x < c \\ R\left(\frac{x-c}{d-c}\right) & \text{if } x \in [c,d] \\ 1 & \text{if } x > d \end{cases} \quad (7)$$

Then, the truth value of *A y's are P* is given as

$$T(A\,y's\,are\,P) = T(A_1\,y's\,are\,P) - T(\overline{A_2}\,y's\,are\,P) \quad (8)$$

For example, if the function $A(x)$ is of form $Trap(a,b,c,d)$, then it is split as $A_1(x) = Trap(a,b,1,1)$ and $\overline{A_2}(x) = Trap(c,d,1,1)$, as illustrated in Figure 2. Note that for quantifiers with monotonic non-increasing membership functions such as *Almost none* in Figure 1, the function $A_1(x)$ is always 1, hence the truth value of *A_1 y's are P* in this case is also 1. Therefore, the truth value of *A y's are P*, is computed as

$$T(A\,y's\,are\,P) = 1 - T(\overline{A_2}\,y's\,are\,P)$$

Also, for monotonic non-decreasing quantifiers like *Many* in Figure 1, the function $A_2(x)$ is always 1. Hence,

T(A y's are P) = T(A_1 y's are P).

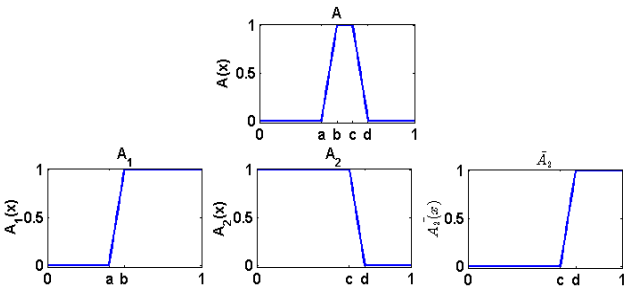

Figure 2: Illustration of how membership functions are split in order to compute truth values

Table 3 shows the truth values of the summaries of the form *A y's are big* computed by this technique with the quantifiers shown in Figure 1. It can be observed that the problem mentioned with the method of [1] does not exist here, and the truth value computed for *Almost None of the balls are big* and *Few balls are big* are more intuitive; *Almost None* being the one with highest truth value for bag 2 and 3. However, the truth value for *Many balls are big* is greater than that of *Few balls are big* and *Some balls are big* in both of these cases, which is counter intuitive according to the data and the semantic order of the quantifiers. Moreover, in bag 1 , even though the truth value of *Many balls are big* is correctly computed to be the highest, that of *Few balls are big* and *Almost None of the balls are big* being greater than *Some balls are big*, is again, not intuitive. We address this problem in the next section.

Table 3: Truth values using [2], of LPS associated with example 1

	Almost None	**Few**	**Some**	**Many**
Bag 1	0.10	0.20	0.00	0.70
Bag 2	0.70	0.10	0.10	0.20
Bag 3	0.70	0.20	0.00	0.30

III. PROPOSED METHOD

The importance of our approach is based on the fact that only one summary may not be sufficient to completely describe a dataset. For example, in [13] all LPS above a threshold were considered to compute linguistic prototypes from a collection of summaries. In such cases, semantically unordered truth values may result in anomalous output. Hence, we lay stress on a technique in which truth values follows the semantic order of the quantifiers. To this end, we modify the method in [2] to produce more intuitive truth values.

For quantifiers A, whose membership function $A(x)$ are of the form defined in equation (1), similar to [2], we split them up into $A_1(x)$ and $A_2(x)$ as shown in equation (6). Then, the truth value of *A y's are P* is given by:

$$T(A\,y's\,are\,P) = T(A_1\,y's\,are\,P) \wedge T(A_2\,y's\,are\,P) \quad (9)$$

where, $A_1(x)$ and $A_2(x)$ are defined as described in Section II.C and the truth values are computed using equation (5). Note that *T(A_2 y's are P)* is evaluated by defining $\overline{A_2}(x)$ as shown in equation (7).

Table 4 shows the truth value of the bags of balls in Table 1 using proposed method. For bags 2 and 3, the truth value for *Almost None of the balls* are big is highest and it varies gradually from *Few balls are big* to *Many balls are big*. Similarly, for bag 1, the truth value is highest for *Many balls are big* and it varies intuitively from *Some balls are big* to *Almost none of the balls are big*. This evidence suggests that the proposed method does not suffer from the discrepancies observed in [1] and [2]. We now show that the truth values computed by this method always follow the semantic order of the quantifiers.

Table 4: Truth value computation using proposed method, of LPS associated with example 1

	Almost None	**Few**	**Some**	**Many**
Bag 1	0.10	0.30	0.30	0.70
Bag 2	0.70	0.40	0.30	0.20
Bag 3	0.70	0.50	0.30	0.30

A. Lemma 1

For any summarizer, P, if there are two monotonic quantifiers A and B such that

$A(x) \geq B(x) \; \forall \; x \in [0,1]$ Then,

$T(A\,y's\,are\;P) \geq T(B\,y's\,are\;P)$, where,

$T_A = T(A\,y's\,are\;P)$, $T_B = T(B\,y's\,are\;P)$ are computed using equation (5).

Proof:

For each α, we have $A(P_\alpha) \geq B(P_\alpha)$. Hence, for any finite set $\alpha_1 \leq \alpha_2 \leq .... \leq \alpha_N = 1$

$$\alpha_1 \wedge A(P_{\alpha_1}) \geq \alpha_1 \wedge B(P_{\alpha_1})$$
$$\alpha_2 \wedge A(P_{\alpha_2}) \geq \alpha_2 \wedge B(P_{\alpha_2})$$
$$...$$
$$...$$
$$\alpha_n \wedge A(P_{\alpha_n}) \geq \alpha_n \wedge B(P_{\alpha_n})$$

Therefore,

$\max_\alpha(\alpha \wedge A(P_\alpha)) \geq \max_\alpha(\alpha \wedge B(P_\alpha))$, and so, $T_A \geq T_B$.

Based on this Lemma, we can prove the main result.

B. Theorem 1

The truth values computed by the proposed method follows the semantic order of the quantifiers. That is, suppose we have three semantically ordered quantifiers *A*, *B* and *C* with their membership functions *A*(*x*), *B*(*x*) and *C*(*x*) respectively, each of the form defined in equation (1), and suppose that

$$T(A\,y's\,are\;P) \geq T(B\,y's\,are\;P). \text{ Then}$$
$$T(B\,y's\,are\;P) \geq T(C\,y's\,are\;P)$$

Proof:

Let, T_A = Truth value of the statement, *A y's are P,* where *A*(*x*) is the membership function of the quantifier *A,* along with similar notations for quantifiers *B* and *C.* Assume that the quantifiers *A*, *B* and *C* are semantically ordered. To compute the truth value as shown in equation (8), the quantifiers are transformed to monotonic functions as shown in equation (6). It is straight forward to see that

$$A_1(x) \geq B_1(x) \geq C_1(x), \text{ and} \tag{10}$$

$$C_2(x) \geq B_2(x) \geq A_2(x) \tag{11}$$

Now using Lemma 1 and equations (10) and (11),

$$T_{A_1} \geq T_{B_1} \geq T_{C_1} \tag{12}$$

$$T_{C_2} \geq T_{B_2} \geq T_{A_2} \tag{13}$$

It is given that $T_A \geq T_B$ (14)

We start by computing T_B,

Case 1, $T_A = T_{A_1}$ **, that is,** $T_{A_2} \geq T_{A_1}$

From equation (12) and (13) we have

$T_{A_1} \geq T_{B_1}$ and $T_{B_2} \geq T_{A_2}$. Hence,

$T_{B_2} \geq T_{A_2} \geq T_{A_1} \geq T_{B_1}$, and so

$T_B = min(T_{B_1}, T_{B_2}) = T_{B_1}$

Case 2, $T_A = T_{A_2}$

Suppose that $T_B = T_{B_2}$,

According to equation (13), $T_{B_2} \geq T_{A_2}$, which would imply that $T_B \geq T_A$. This contradicts the hypothesis of the theorem, restated in equation (14). Hence, if the given condition holds then irrespective of which component defines the truth value T_A,

$$T_B = \min(T_{B_1}, T_{B_2}) = T_{B_1} \tag{15}$$

Similarly, to compute T_C, from equations (12), (13) and (15) respectively we have,

$T_{B_1} \geq T_{C_1}$, $T_{C_2} \geq T_{B_2}$ *and* $T_{B_2} \geq T_{B_1}$. Thus,

$T_{C_2} \geq T_{B_2} \geq T_{B_1} \geq T_{C_1}$, which implies that

$T_C = \min(T_{C_1}, T_{C_2}) = T_{C_1}$.

Now, since $T_{B_1} \geq T_{C_1}$ and that $T_B = T_{B_1}$ and $T_C = T_{C_1}$, we conclude that,

$T_B \geq T_C$. That is, $T(B\,y's\,are\;P) \geq T(C\,y's\,are\;P)$ and the proof is complete.

From this theorem, it's clear that the truth values will follow the semantic ordering in all directions along with the quantifiers.

IV. CASE OF EXTENDED PROTOFORMS

Linguistic summaries with simple protoforms: *A y's are P* described above have been extended to the protoforms: *A y's are P and Q* and *A R y's are P* in the past. They can be exemplified by *Some of the balls are big and heavy* and *Some of the big balls are heavy*, respectively. As a naming

convention, we call summaries of the form *A y's are P and Q* as protoform 2 summaries and *A R y's are P* as protoform 3 summaries. Both of these extended summaries encapsulate more than one feature of data at the same time, hence containing richer information as compared to simple protoforms

Along with the truth value, the protoform 3 summaries are accompanied by another metric called the degree of focus [9]. The degree of focus conveys information about how applicable is the summary with respect to *R* and also enables to discard non-promising summaries. Similar to the case of simple protoforms, we observe a discrepancy when the truth values are computed using Zadeh calculus of linguistically quantified propositions [25] and extend our method to produce more intuitive truth values.

A. *Truth value computation using Zadeh calculus of linguistically quantified propositions*

Equation (16) and (17) show the method used to compute the truth values of extended protoforms using Zadeh calculus with the formula for degree of focus shown in equation (18)

$$T(A\ y's\ are\ P\ and\ Q) = A\left(\frac{1}{N}\sum_{i=1}^{N}(P(y_i)\wedge Q(y_i))\right) \tag{16}$$

$$T(A\ R\ y's\ are\ P) = A\left(\frac{\sum_{i=1}^{N}(P(y_i)\wedge R(y_i))}{\sum_{i=1}^{n}R(y_i)}\right) \tag{17}$$

$$d_f(A\ R\ y's\ are\ P) = \frac{1}{N}\sum_{i=1}^{N}R(y_i) \tag{18}$$

where, similar to the simple protforms, $P(x)$ is the membership function of the summarizer, $A(x)$ is the membership function of the quantifier and N the size of the set $\{y_i \mid i = 1,2...N\}$. For protoform 3 summaries, R is called the qualifier with its membership function represented by $R(x)$ and for the case of protoform 2 summaries, Q is another summarizer along with P. Also, it is worth noting that for protoform 2 summaries, the degree of focus is not applicable since it does not add any information.

In the following, we present some examples to illustrate the discrepancy when the truth values are computed by equation (16) and (17) and show that our method solves the mentioned problem. Table 5 shows the bigness and heaviness of 4 bags of balls containing 10 balls each. We wish to compute truth values of the extended protoform summaries of the form: *A of the balls are big and heavy* and *A of the big balls are heavy*, *A* being each of the quantifiers, *Almost None*, *Few*, *Some* and *Many* as defined in Figure 1 in Section II.

From the data presented in the Table 5, intuitively, bag 1 is best described by *Few of the big balls are heavy* and *Few of the balls are big and heavy* while bag 2 by *Almost None of the big balls are heavy* and *Almost none of the balls are big and heavy*. To point out that both of the extended summaries conveys

Table 5: Bigness and Heaviness of 4 bags of balls

Bag1	Bigness	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
	Heaviness	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bag2	Bigness	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
	Heaviness	0.3	0.3	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0
Bag3	Bigness	0.9	0.8	0.8	0.8	0.0	0.0	0.0	0.0	0.0	0.0
	Heaviness	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bag4	Bigness	0.9	0.9	0.9	0.9	0.3	0.3	0.3	0.3	0.3	0.3
	Heaviness	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0

similar but different type of information about the data, we designed data for bag 3 and 4, such that bag 3 is best represented by *Almost None of the balls are big and heavy* (since there is only one big ball big that is heavy) and *Few of the big balls are heavy* (out of four really big balls, one is heavy). Similarly, for bag 4, the summary *Many of the big balls are heavy* describes it well, as all the four really big balls have higher memberships in heaviness. However, due to the fact only four balls are really big and heavy, the summary *Some of the balls are big and heavy* is equally true but conveys different information.

Tables 6 and 7 shows the truth values computed using equation (16) and (17), respectively, with the degree of focus shown for protoform 3 summaries in the last column of Table 7. We observe that for bags 1, 3 and 4, this method computes the highest truth values for both the protoform 2 and protoform 3 summaries that best represents the data. However, in case of bag 2, the truth value of *Few of the big balls are heavy* and *Few of the balls are big and heavy* being highest does not look intuitive, since none of the balls are really heavy (their membership being 0.3). Similarly, for bag 4, even though the truth value of *Some of the balls are big and heavy* is highest, it being very close to *Few of the balls are big and heavy* is non intuitive. This discrepancy arises due to the use of average operator similar to the case of simple protoforms. We also note,

Table 6: Truth values of summaries of the form *Almost none of the balls are big and heavy, Few of the balls are big and heavy etc.* computed using the Zadeh calculus of quantified propositions

	Almost None	**Few**	**Some**	**Many**
Bag 1	0.00	1.00	0.00	0.00
Bag 2	0.50	0.75	0.00	0.00
Bag 3	1.00	0.45	0.00	0.00
Bag 4	0.00	0.40	0.60	0.00

Table 7: Truth values and degree of focus of summaries of the form *Almost none of the big balls are heavy, Few of the big balls are heavy etc.* computed using the Zadeh calculus of quantified propositions

	Almost None	**Few**	**Some**	**Many**	**dFoc**
Bag 1	0.00	0.70	0.30	0.00	0.60
Bag 2	0.00	1.00	0.00	0.00	0.50
Bag 3	0.00	1.00	0.00	0.00	0.33
Bag 4	0.00	0.00	0.00	1.00	0.54

none of the summaries can be rejected on the basis of the degree of focus, since it is fairly high for all of the 4 bags.

B. Proposed method adapted to the extended protoforms

We modify the method proposed earlier in this paper to compute truth values of the extended summaries with non-decreasing quantifiers, as shown in equation (19) and (20).

$$T(A\ y's\ are\ P\ and\ Q) = \max_{\alpha}\left(\alpha \wedge A\left(P_{\alpha}^{Q_\alpha}\right)\right) \tag{19}$$

where,

$$Q_\alpha = \{y_i \in Y \mid Q(y_i) \geq \alpha\},\ P_{\alpha}^{Q_\alpha} = \frac{\left|\{y_i \in Q_\alpha \mid P(y_i) \geq \alpha\}\right|}{N}$$

$$T(A\ R\ y's\ are P) = \max_{\beta \in [0, max(R(y_i))]} \beta \wedge \left(\max_{\alpha \in [0,1]} \left(\alpha \wedge A(P_{\alpha}^{R_\beta}) \right) \right) \tag{20}$$

$$d_f^{R_\beta} = \frac{1}{N}\sum_{i=1}^{|R_\beta|} R(y_i)\ such\ that\ y_i \in R_\beta \tag{21}$$

where, $R_\beta = \{y_i \in Y \mid R(y_i) \geq \beta\}$,

$P_{\alpha}^{R_\beta} = \frac{\left|\{y_i \in R_\beta \mid P(y_i) \geq \alpha\}\right|}{\left|R_\beta\right|}$, for $\left|R_\beta\right| > 0$.

We note that for cases where $\left|R_\beta\right| = 0$, the truth value of the summary for that β cut is set to 0, since there are no objects to be summarized.

To elaborate, for protoform 2 summaries, we modify equation (5) by implementing the 'and' between summarizers with a minimum operator. That is, in equation (19), for each object we take the smaller of the memberships in *P* and *Q*, and then compute the truth value of this new set.

For the case of protoform 3 summaries, we take beta cuts of the qualifier data from 0 to the highest membership of the objects in the qualifier *R* (i.e., $max(R(y_i))$). For each of the objects falling in the current beta cut, we follow the procedure for the simple protoforms. The intuition behind this is that when computing truth values of summaries like *Some of the big balls are heavy*, we should only focus on the balls that are *big* under certain condition, which is the beta cut in equation (20). Next, as shown in equation (21), the degree of focus is computed for the beta cut that produces this truth value, that is, we compute the proportion of objects that have the memberships in the qualifier, *R*, greater than the value of beta that resulted in the truth value in equation (20). This results in a degree of focus for each summary, unlike the Zadeh calculus method, where there is one degree of focus for each dataset to be summarized. Also, for summaries with zero truth value, the degree of focus is not applicable since it does not provide any information.

For quantifiers whose membership functions are not monotonic non-decreasing, similar to the simple protoforms, we split the quantifiers as shown in section II.C and use equation (8) to compute the final truth value.

Tables 8 and 9 shows the truth value of the summaries of four bags shown in Table 5, computed using equation (19) and (20) respectively. For bag 2, *Almost none of the big balls are heavy* has the highest truth value with the same degree of focus for all four summaries. In the case of bag 4, the truth value of *Many of the big balls are heavy* is computed to be highest with a degree of focus of 0.36. For this bag, *Some of the big balls are heavy* has a high degree of focus, but it can be rejected on the basis of a very low truth value. It is also worth noting that for bag 3, the truth value of *Many of the big balls are heavy* is computed to be maximum, however, with a very low degree of focus. This is due to the fact that the only really heavy ball has bigness of 0.9. Hence, this summary can be interpreted as *Many balls with bigness above 0.9 are heavy* (which seems to be true). But we can discard this summary on the grounds of having very small degree of focus. For the same bag, the truth value of *Few of the big balls are heavy* is computed to be 0.8 with a comparatively higher degree of focus of 0.33, which seems to be the correct representation of the balls in bag 3.

For the case of protoform 2 summaries, our method produces expected truth values for all four of the bags. Unlike the method involving the Zadeh calculus, we end up with *Almost None of the balls are big and heavy* as the summary with highest truth value for bag 2 which is in correspondence with the data. Also, for bag 4, *Some of the balls are big and heavy* has a very high truth value in comparison to *Few of the balls are big and heavy* which looks more correct intuitively.

Table 8: Truth values of summaries of the form *Almost none of the balls are big and heavy, Few of the balls are big and heavy etc.* computed using proposed method

	Almost None	**Few**	**Some**	**Many**
Bag 1	0.00	1.00	0.00	0.00
Bag 2	0.70	0.30	0.30	0.00
Bag 3	1.00	0.50	0.00	0.00
Bag 4	0.10	0.10	0.90	0.00

Table 9: Truth values and degree of focus of summaries of the form *Almost none of the big balls are heavy, Few of the big balls are heavy etc.* computed using proposed method

	Almost None		**Few**		**Some**		**Many**	
	T	**dFoc**	**T**	**dFoc**	**T**	**dFoc**	**T**	**dFoc**
Bag 1	0.00	NA	0.70	0.60	0.30	0.60	0.00	NA
Bag 2	0.70	0.50	0.30	0.50	0.30	0.50	0.30	0.50
Bag 3	0.00	NA	0.80	0.33	0.00	NA	0.90	0.09
Bag 4	0.00	NA	0.00	NA	0.20	0.54	0.90	0.36

V. Conclusions

In this work we identified some discrepancies in previous techniques to compute truth values of linguistic protoform summaries of the form *Few balls are big*, *Few of the balls are big and heavy* and *Few of the big balls are heavy*. We solve the problems associated with these methods and show that our solution computes intuitive truth values of linguistic summaries with both simple and extended protoforms. Moreover, we provide a proof to show that our method always produces truth values according to the semantic order of language it represents. Linguistic protoform summaries have been used in numerous applications in the past. A fresh look at them with our new method might provide important observations.

Acknowledgment

This research is supported in part by Center for Eldercare and Rehabilitation Technology at the University of Missouri.

References

[1] R. R. Yager, "A new approach to the summarization of data," *Information Sciences,* vol. 28, pp. 69-86, 1982.

[2] A. Wilbik, U. Kaymak, J. Keller, and M. Popescu, "Evaluation of the Truth Value of Linguistic Summaries – Case with Non-monotonic Quantifiers," in *Intelligent Systems'2014*. vol. 322, P. Angelov, K. T. Atanassov, L. Doukovska, M. Hadjiski, V. Jotsov, J. Kacprzyk*, et al.*, Eds., ed: Springer International Publishing, 2015, pp. 69-79.

[3] IBM. *The Four V's of Big Data*. Available: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[4] A. Gatt, F. Portet, E. Reiter, J. Hunter, S. Mahamood, W. Moncur, and S. Sripada, "From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management," *Ai Communications,* vol. 22, pp. 153-186, 2009.

[5] A. Wilbik, J. M. Keller, and G. L. Alexander, "Linguistic summarization of sensor data for eldercare," in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, 2011, pp. 2595-2599.

[6] E. Reiter and R. Dale, "Building Natural Language Generation Systems," 2006.

[7] D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, and M. Aud, "Linguistic summarization of video for fall detection using voxel person and fuzzy logic," *Computer Vision and Image Understanding,* vol. 113, pp. 80-89, 2009.

[8] R. Castillo-Ortega, N. Marín, and D. Sánchez, "Time series comparison using linguistic fuzzy techniques," in *Computational Intelligence for Knowledge-Based Systems Design*, ed: Springer, 2010, pp. 330-339.

[9] J. Kacprzyk and A. Wilbik, "Towards an efficient generation of linguistic summaries of time series using a degree of focus," in *Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American*, 2009, pp. 1-6.

[10] J. Kacprzyk, A. Wilbik, and S. Zadrozny, "Linguistic Summaries of Time Series via a Quantifier Based Aggregation Using the Sugeno Integral," in *FUZZ-IEEE*, 2006, pp. 713-719.

[11] J. Kacprzyk, A. Wilbik, and S. Zadrożny, "Linguistic summarization of time series using a fuzzy quantifier driven aggregation," *Fuzzy Sets and Systems,* vol. 159, pp. 1485-1499, 2008.

[12] J. Kacprzyk, A. Wilbik, and S. Zadrożny, "An approach to the linguistic summarization of time series using a fuzzy quantifier driven aggregation," *International Journal of Intelligent Systems,* vol. 25, pp. 411-439, 2010.

[13] A. Wilbik, J. M. Keller, and J. C. Bezdek, "Linguistic Prototypes for Data From Eldercare Residents," *Fuzzy Systems, IEEE Transactions on,* vol. 22, pp. 110-123, 2014.

[14] B. Bouchon-Meunier and G. Moyse, "Fuzzy linguistic summaries: Where are we, where can we go?," in *Computational Intelligence for Financial Engineering & Economics (CIFEr), 2012 IEEE Conference on*, 2012, pp. 1-8.

[15] M. Delgado, M. D. Ruiz, D. Sánchez, and M. A. Vila, "Fuzzy quantification: a state of the art," *Fuzzy Sets and Systems,* vol. 242, pp. 1-30, 2014.

[16] M. Sugeno, "Theory of fuzzy integrals and its applications," Tokyo Institute of Technology, 1974.

[17] R. Castillo-Ortega, N. Marin, and D. Sánchez, "Linguistic local change comparison of time series," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, 2011, pp. 2909-2915.

[18] A. Wilbik and J. M. Keller, "A distance metric for a space of linguistic summaries," *Fuzzy Sets and Systems,* vol. 208, pp. 79-94, 2012.

[19] A. Wilbik and J. M. Keller, "Anomaly detection from linguistic summaries," in *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, 2013, pp. 1-7.

[20] G. Trivino and M. Sugeno, "Towards linguistic descriptions of phenomena," *International Journal of Approximate Reasoning,* vol. 54, pp. 22-34, 2013.

[21] L. Liétard, "A functional interpretation of linguistic summaries of data," *Information Sciences,* vol. 188, pp. 1-16, 2012.

[22] R. Castillo-Ortega, N. Marín, D. Sánchez, and A. G. Tettamanzi, "Linguistic summarization of time series data using genetic algorithms," in *EUSFLAT*, 2011, pp. 416-423.

[23] C. Donis-Díaz, A. Muro, R. Bello-Pérez, and E. V. Morales, "A hybrid model of genetic algorithm with local search to discover linguistic data summaries from creep data," *Expert Systems with Applications,* vol. 41, pp. 2035-2042, 2014.

[24] A. Kaufmann and M. M. Gupta, *Introduction to Fuzzy Arithmetic: Theory and Applications*: Van Nostrand Reinhold Company, 1985.

[25] L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy sets and systems,* vol. 90, pp. 111-127, 1997.