Preliminary Study for a Survey-Based Fuzzy Membership Function Definition for Imprecise Quantification in Croatian*

Leo Mršić, Ph.D.
Algebra University College
Zagreb, Croatia
leo.mrsic@algebra.hr

Sandro Skansi, Ph.D.
University of Zagreb
Zagreb, Croatia
skansi.sandro@gmail.com

Robert Kopal, Ph.D.
Algebra University College
Zagreb, Croatia
robert.kopal@algebra.hr

Abstract. *In this preliminary report we propose a survey-based method for defining imprecise quantification for Croatian. By using the results of a survey conducted among students, a fuzzy membership function for each of the precise and imprecise quantification terms can be defined, with possible extensions to type-2 fuzzy memberships. An earlier version of this paper was submitted and subsequently withdrawn from the ACE-X 2017 conference.*

Keywords. Fuzzy Set Membership, Fuzzy Quantifiers, Linguistic Variables, Croatian Quantification

1 Introduction

Quantifiers in language and linguistic variables in fuzzy logic share a common theme. The study of quantification dates back to at least Aristotle's Organon, and in the modern period was revived by G. Frege (Frege 1879) and perfected by C. S. Peirce (Peirce 1885). Quantifiers today are an essential feature of all major applied logic systems with a few noteworthy exceptions (SAT-related logics (Marek 2009), propositional modal logic (Blackburn, de Rijke, Venema 2002)). Quantifiers however, are an essentially linguistic constructs allowing the reference to a number of terms (Peters and Westerstahl 2006). Unprecise quantifiers were a motivational factor behind the development of mathematical fuzzy logic (Hajek 1998), and a natural approach to their meaning is via fuzzy membership functions. Linguistic variables were at first considered in the context of process theory (D'Ambrosio 1989), but as fuzzy logic became more know today they are discussed in the context of fuzzy logic (Ross 2010) (Mrsic 2017).

A second interesting phenomenon is that nonconventional quantifiers (Torza 2015) (Peters and Westerstahl 2006) are relative to a given language, and different languages possess natural quantifier terms both for precise and imprecise numberings. Precise quantifiers can be thought of as number terms, uniquely denoting a precise quantity. An example of such term could be "ten", but their apparent precision fails to take in account their ability to reference an entity (Donnellan 1966) (Donnellan 1972). A statement "Fetch me that jar with ten bolts" may be successful at referring to the correct jar (containing e. g. 12 bolts). The example might be made even more elaborate by stipulating the presence of another jar with e.g. 143 bolts next to it. This may seem like a minor point, but it points out to the inherent imprecision even with precise numbering terms when considering the everyday communicational aspects of the language quantifiers.

(A first version of this paper under the title "Learning Fuzzy Membership functions for Slavic Quantifier Terms" was intended to be published with the ACE-X 2017 conference, but we felt that the paper at that time needed substantial revision and expansion, as well as refocusing, and we have subsequently withdrawn it from publication prior to the conference. The present paper is a revised and refocused version of the previous unpublished and unpresented paper.)

2 Basic quantifier and number term usage

The basic quantifiers in first order logic are "all" and "exists", and they are defined to be true when combined with a property that holds. An example of this could be "For all x, P(x)". Naïve set theory, prior to Russell's paradox (van Heijenoort 1967) claimed that every property P(*) defines a set of objects x satisfying P(*). Classical logic together with naïve set theory was proven inconsistent by Russell's paradox. It is an open question whether naïve set theory and fuzzy logic is inconsistent (Behounek and Hanikova 2014), but we will assume it is for the scope of this paper, since it simplifies the exposition. The quantifier "Exists" holds true if the property holds for at least one object. One would think quantifiers like "Ten" could be easily defined by an extension of this principle, but it then we run into the problem of reference. If we define "Ten" to mean exactly 10, the reference should have failed for the jar with ten bolts.

* This paper is published and available in Croatian language at: http://ceciis.foi.hr

It is a point of fact that reference in such cases succeeds and the solution is to model the desired quantifier terms with fuzzy memberships. Let us give an example. As we stated earlier, "There exists an x such that P(x)" is true provided there is an object with the property P(*). We could make the same analogy for "Ten" then it would be that we have to find a truth value for "There are ten x such that P(*)". We could say it is true when there are ten items, and false otherwise, but we could also relax this condition and accept some border cases with some truth, e. g. with a "truthness" of say 0.8 for 9 and 11 items.

The terms analyzed were (i) the precise terms and (ii) the non-precise terms. The precise terms analyzed were "Jedan", "Dva", "Tri", "Četiri", "Pet", "Šest", "Sedam", "Osam", "Devet", "Deset", "Jedanaest","Dvanaest", "Trinaest" (numbers one to thirteen). The non-precise terms were "Jedva išta" (barely anything), "Par" (a couple), "Nekolicina" (few), "Nekoliko" (few), "Brojni" (numerous), "Dosta" (plenty), "Mnogo" (a lot), "Puno" (a lot), "Malo" (few), "Nešto" (some), "Osjetno" (quite a few), "Više" (a number of). As it can be seen we have included several different words but with very similar meaning, so the distinction between their membership functions can be seen as a contribution to the understanding of their semantics. Their translation in English is also tentative at best, as many of them are considered synonyms.

For finding a good represenation of the linguistic quantifiers, we have conducted a survey among 43 students of the University College Algebra asking them to fill in a chart with scores 1-5 representeting how good a term describes a quantity. For example, a score of 4 for "Mnogo" under the column "Quantity: 20" meant that "Mnogo" was 80% appropriate term for describing a quantity of 20 units. The 1-5 scores were subsequently normalized to a 0-1 scale.

A more comprehensive survey, as well as the interpretation of results as type-2 fuzzy sets is planned for further research. Another topis for further research is the use of our approach to facilitate anaphora resolution in South Slavic languages.

2.1 Precise numbering terms

It turns out that our surveys reported such fuzziness even for precise numbers, and we interpolated values to the results to find a function to describe the graph. The most common number terms in Croatian were considered and most of them showed a relatively high precision, along with some fuzziness on the border cases, which increased as the numbers increased. The notable exception was "Deset", or (ten) showing considerably more fuzziness than "Jedanaest" (eleven).

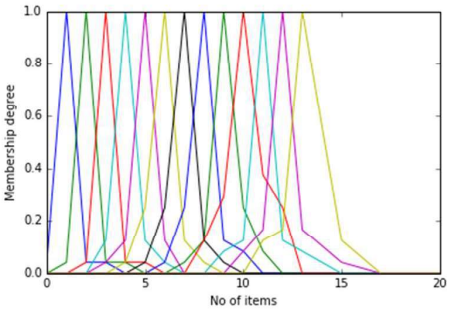

Figure 1. Precise numbering terms

All of these functions can be modelled by a minimal adjustment to the Gaussian function, or, for our needs, with a simple piecewise spike function. Eventual gaps in values are inconsequential, as long as the function is defined for all arguments. Also, the function should be used to assess only integer values.

2.2 Non-Precise numbering terms

The first non-precise quantifier term analyzed was "Nekolicina" (few), which displayed a bell-like shape and could easily be approximated by a Gaussian function. The visible hops on the right-hand side do not matter much for modelling, but the bounds do, so the function returns 0 on 0, 0.7 on 5, and again goes to 0 at 12. The graph is shown on the picture below.

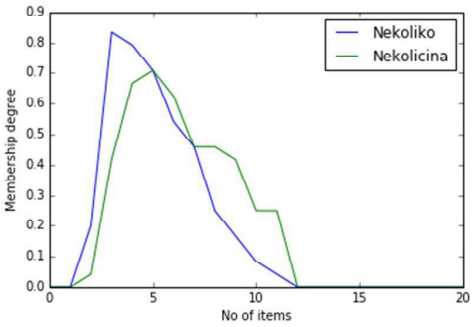

Figure 2. Non-Precise numbering terms

The second nonprecise term we modelled was "Nekoliko" (few). The first part of the function acts similarly to the previous one, but topping off at 0.8 when 3 is reached, and going down more steeply at 12. Notice how the values at 10 differ considerably. The most important difference is the argument for which the functions achieves it maximal value (in the previous case 5, and here 3), which points to the semantic differences in these two terms (both adequately translated into English with the word "few").

This points to the fact (which we shall see soon to hold) that for most quantifier term membership

functions, the argmax, argmin, maximal and minimal values give a very precise definition for a fuzzy membership function. We could approach this issue by using a Gaussian function to model them, but we shall use a piece-wise linear function instead (it is computationally more feasible).

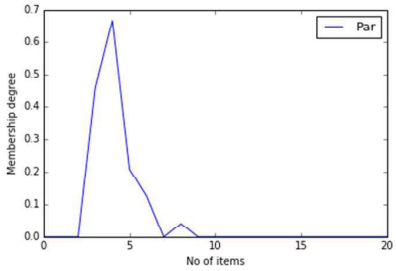

Figure 3. Word "Par" graph

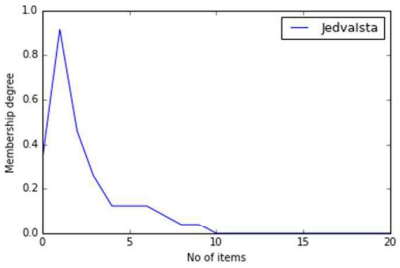

Figure 4. Word "Jedva išta" graph

The word "Par" (a couple), achieves its maximum at 4, with a very steep slope (ending at 6), but in practice it is the same as "Nekoliko", only with the maximum translated at 4, and a steeper curve. "Jedva išta" (barely anything) displays a similar pattern, with a high maximum at 0.95 and argmax at 1.

The next type of membership functions are ReLU-type functions, which are similar to appearance to the ReLU function (f(x)=max(0,x)). They are large quantifier terms, and they are dependent on scale: if the scale goes up to 1000, then at 1000 they will reach the value closest to 1. If on the other hand they are scaled up to 10000, then there the membership functions will be close to 1.

The term "Dosta" (plenty) is the most peculiar, starting the rise at 5, spiking at 10 and plateauing at 10-17, and taking off afterwards. This indicates the problem with the scale limit, and it can be restated to catch the fact that "Dosta" has a spike in mid-range, a max at the end of the range and a plateau between.

"Puno" and "Mnogo" (a lot) are sometimes considered synonyms in south Slavic languages, but they seem to different in semantics, and shown on the images below, in since "Mnogo" was assessed truer of a smaller number of items than "Puno".

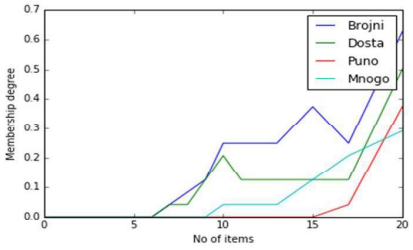

Figure 5. Comparison graph

The step in "Mnogo" is an interesting occurrence but it is inconsequential. The semantic difference is really a matter of nuances in this case, and there is no practical applicability of this, but it is an interesting curiosity nevertheless.

The term "Brojni" (numerous), has a more erratic membership function shown on the image below, but can still be approximated by a ReLU-like function. A detailed exposition of how the relevant ReLU's are defined is given in the next section.

The terms "Malo" (few) and "Nešto" (some) have a spike at 2 and 3, to decline later on. They are similar in behaviour to number terms, only with a prolonged tail as the values grow.

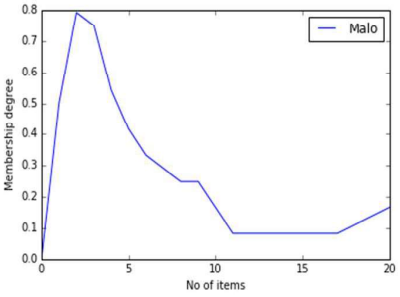

Figure 6. Word "Malo" graph

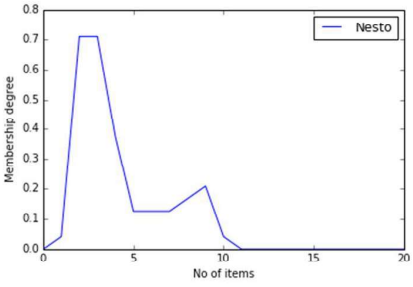

Figure 7. Word "Nešto" graph

The last two term are "Osjetno" (considerably) and "Više" (more), which share a middle spike at 10, and a subsequent decline, followed by a rise to 1 as the

number of counted terms rise. This is in a way surprising, since their semantics is traditionally not considered close, whereas "Više" is considered almost synonymous with "Puno" and "Mnogo". For the purpose of extracting membership functions however, we shall regard "Više" as being a ReLU-like function.

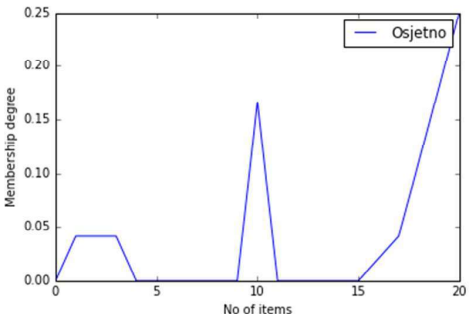

Figure 8. Word "Osjetno" graph

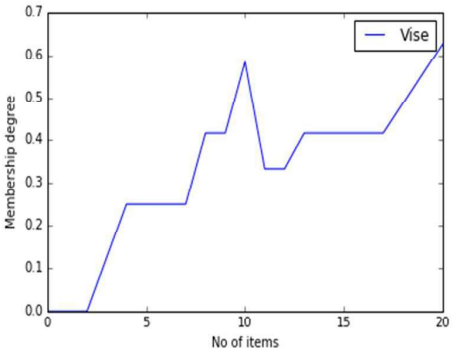

Figure 9. Word "Više" graph

3 Extrapolated functions

We give a table of the extrapolated functions:

Table 1. Extrapolated functions: translations, Non-Zero range

Term	Translation	Non-Zero range
Nekoliko	Few	1-12
Nekolicina	Few	1-12
Par	A couple	2-6
Jedva išta	Barely Anything	0-10
Brojni	Numerous	6-Inf(20)
Dosta	Plenty	6-Inf(20)
Puno	A lot	15-Inf(20)
Mnogo	A lot	9-Inf(20)
Više	A number of	2-Inf(20)
Malo	Few	0-20
Nešto	Few	1-11
Osjetno	Considerably	0-4, 9-11, 15-Inf(20)
[Number N]	Precise number N	(N-1)-(N+2)

Table 2. Extrapolated functions: Maximum Value At, MF type, ReLU kickoff

Term	Maximum Value At (Argument, Value)	MF type	ReLU kickoff
Nekoliko	(3, 0.85)	Spike	--
Nekolicina	(5, 0.7)	Spike	--
Par	(4, 0.65)	Spike	--
Jedva išta	(1, 0.9)	Spike	--
Brojni	(20, 0.65)	ReLU	0.33
Dosta	(20, 0.5)	ReLU	0.33
Puno	(20, 0.38)	ReLU	0.75
Mnogo	(20, 0.3)	ReLU	0.45
Više	(20, 0.6)	ReLU	0.1
Malo	(2, 0.8)	Spike	--
Nešto	(3, 0.7)	Spike	--
Osjetno	(2, 0.05), (10, 0.16), (20, 0.25)	Other	--
[Number N]	(N, 1.0)	Spike	--

To extract the exact version of the membership function needed from the above table, the membership function type column must be consulted first. In the case of ReLU functions, we use the following general form:

$$ReLU(x) = max(0, f(x)) \quad (1)$$

Where f(x) is a linear function calculated through two points after the beginning of the non-zero range. This is a trivial task, but we repeat the process for the reader's convenience (for details see Bronshtein et al. 2007). First the slope is calculated through two points of the non-zero range with slope(f) = (f(x2)-f(x1))/(x2-x1), where the usual denominator provisions are enforced (but even if not, this is not consequential since a different pair of points is chosen and the slope is calculated). After the slope is calculated, by using y-y1 = slope(f)(x2-x1), an explicit representation is obtained. A word of advice: when choosing points for a calculation, due to approximation errors is best to choose the points at which the piece-wise function connects with its other parts. This means that when calculating the slope, the last point for which f(x) = 0 should be one of the points used, and the second should be the end of the range point (which has also the

maximal value in ReLU-like functions). This is even more important in the spike functions, which have a general form:

$$Spike(x) = max(0, (up(x), down(x))) \quad (2)$$

Where the (up, inflection, down) is a shorthand for a two-piece function with the (global) maximum in the middle (this is the inflection point). Two slopes are needed. The first slope is the "up" portion. Its calculation is similar to ReLU's. Two points are used, the left one being the last point for which f(x)=0, which is the first point in the non-zero range from the table above, and for the second point the pair in the column Maximal Value At from the table above should be consulted. For the "down" part, the first point for the slope should be the pair in the column Maximal Value At from the table above, and the second point the first point for which f(x)=0 after the inflection point.

4 Conclusion

In our research we have focused more on smaller number terms, due to the fact that they are effectively bounded by 0. These terms are the spike-type functions. For the sake of completeness, we have also included the large quantifier terms represented by the ReLU-like membership functions, but there is an inherent problem with these terms, namely that they are interpreted relative to the scale offered: In some cases, 20 may be "a lot", and in some cases 100 may not fit "a lot" well. We used a range from 0 to 20, but to provide for this problem we have added a "ReLU kickoff" parameter in the table above, which defines after which percentage of the range the function stops being zero and takes off. This is better than the approach with log scales, as the log scale would require nonlinear function in the ReLU, and yet they would only reduce the problem and not eliminate it completely, since a right-hand bound would still have to be provided.

The membership function for "Osjetno" was left out, as it did not give clear readings (it has a max at only 25%), and it is rather difficult to describe. We feel that to make a useable representation for "Osjetno", a larger dataset and a deeper polynomial fitting algorithm should be used. In this way, more regularities might arise, so we leave this as an open problem for further research. There are of course a number of other terms just as complex as "Osjetno" which still have to be addressed.

We believe our research could be of great use for computational semantics and anaphora resolution in south Slavic languages. First, due to similarities, we believe that for workable engineering applications, the current representations of quantifier terms in Croatian could be used for all the south Slavic languages (Bosnian, Bulgarian, Croatian, Macedonian, Montenegrin, Serbian, Slovene). The application to computational semantics is quite straightforward: the two main segments of computational semantics are the relationships (extracted usually by machine learning) and the quantifiers, which we have partly addressed.

An additional application would be to be able to assess the anaphoric behaviour of quantifier terms, i.e. does a current unspecified quantity refer to a previous unspecified quantity and, if yes, to which one if more than one is mentioned. This can be done by learning how similar are two membership functions given a hypothesized quantity. This can no longer be done with a simple linear fitting, and we leave this for further research.

References

Behounek, L. and Hanikova, Z. (2014). Set Theory and Arithmetic in Fuzzy Logic. In Petr Hajek on Mathematical Fuzzy Logic, ed. F. Montagna, pp. 63-89.

Blackburn, P., de Rijke, M., Venema, Y. (2002). Modal Logic (Cambridge Tracts in Theoretical Computer Science). Cambridge: Cambridge University Press.

Bronshtein, I. N., Semendyayev, K. A., Musiol, G. and Muehlig, H. (2007). Handbook of Mathematics (Fifth Edition). Berlin: Springer.

D'Ambrosio, B. (1989). Qualitative Process Theory Using Linguistic Variables. Berlin: Springer.

Donnellan, K. S. (1966). Reference and Definite Descriptions. The Philosophical Review, vol. 75 No. 3, pp. 281–304.

Donnellan, K. S. (1972). Proper Names and Identifying Descriptions. U D. Davidson i G. Harman (ur.). Semantics of Natural Language.

Frege, G. (1879). Begriffsschrift: eine der arithmetischen nachgebildete Formelsprache des reinen Denkens. Halle.

Hajek, P. (1998). Metamathematics of Fuzzy Logic. Amsterdam: Kluwer Academic Press.

van Heijenoort, J. (1967). From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931, pp. 124-125. Cambridge: Harvard University Press.

Marek, V. W. (2009). Introduction to Mathematics of Satisfiability. Boca Raton, FL: Chapman & Hall/CRC Studies in Informatics Series.

Mrsic, L. and Klepac, G and Kopal R (2017). A New Paradigm in Fraud Detection Modeling Using Predictive Models, Fuzzy Expert Systems, Social Network Analysis, and Unstructured Data, Computational Intelligence Applications in Business Intelligence and Big Data Analytics, Auerbach Publications, pp. 157-194

Peirce, C. S. (1885). "On the Algebra of Logic: A Contribution to the Philosophy of Notation, American Journal of Mathematics, vol. 7, pp. 180–202.

Peters, S. and Westerstahl, D. (2006). Quantifiers in Language and Logic. Oxford: Oxford University Press

Ross, T. (2010). Fuzzy Logic with Engineering Applications. New York: Wiley Press.

Torza. A. (2015). Quantifiers, Quantifiers, Quantifiers: Themes in Logic, Mataphysics and Language. New York: Springer.