Interpretation and Processing Time of Generalized Quantifiers: Why your Mental Space Matters

Alice Ping Ping Tse (tse@tf.uni-freiburg.de) and Marco Ragni (ragni@tf.uni-freiburg.de)
Institut für Informatik, Technische Fakultät, Albert-Ludwigs-Universität Freiburg
Georges-Köhler-Allee, Geb. 052, 79110 Freiburg im Breisgau, Germany

Abstract

Classical quantifiers (e.g., "all", "some" and "none") have been extensively studied in logic and psychology. In contrast, generalized quantifiers (e.g., "most") allow for fine-grained statements about quantities. The discrepancy in the underlying mental representation and its interpretation among interpreters can affect language use and reasoning. We investigated the effect of quantifier type, quantification space (set size) and monotonicity on processing difficulty (in response time, RT) and response diversity of 77 generalized quantifiers. Shannon entropy was employed to measure response diversity. Our findings indicate: (i) Set size is a significant factor of response diversity, which implies that the underlying space is relevant for the interpretation. (ii) Quantifiers possess a rather static underlying representation within and across tasks within a participant. (iii) Quantifier type and monotonicity can affect response diversity; while the response diversity can predict RT. (iv) In reasoning, the number of generalized quantifiers versus classical quantifiers in a syllogism is a factor of response diversity. Diversity in the interpretation of generalized quantifiers may be a cause of human's deviation from logical responses.

Keywords: generalized quantifiers; syllogism; total set size; monotonicity, individual differences

Introduction

"Quantifiers" can have two definitions: In logic, a quantifier acts as the binder to denote the relationships between sets. In natural language, a quantifier is a determiner or pronoun indicative of quantity or amount. In daily English, it limits and modifies the quantity of the noun it is attached to. They map categories to types. Hence, they are the basis for many fundamental concepts in different fields, especially logic, linguistics and psychology. In first-order logic, there are only two basic quantifiers: the universal "for all, ∀" and the existential quantifier "there exist (or for some), ∃", which denote quantities. In Aristotelian logic (Austin, et al., 1971; Westerståhl, 2011), there are three quantifiers, "all/every", "some" (also for "some… not"), and "no". However, the aforementioned first-order quantifiers are too restricted in daily language use. Generalized quantifiers (also known as the second-order predicates or binary quantifiers) are in a wider use in language, for example, when the exact amount is not available (which is quite usual in daily situations) or to emphasize a rather qualitative property of the amount (e.g., "more than half", "most", "a few"). Generalized quantifiers (or just quantifiers) include words and phrases like 'most', 'many', 'few', 'a few', 'some', 'more than half', 'commonly', 'typically' and cardinal numbers (e.g., more than one, and exact numbers such as two, a hundred).

Since the first articles by Barwise and Cooper (1981) in the field of linguistics and Lindström (1966) in the field of logic, an increasing number of research articles have focused on generalized quantifiers. The interpretation of generalized quantifiers can be affected by factors like the quantification space – its total set size (Newstead, Pollard, & Riezebos, 1987), word frequency (Chase, 1969), monotonicity (Szymanik & Zajenkowski, 2013), common belief and background knowledge (Newstead & Collis, 1987; Moxey & Sanford, 1993), and context and working memory (Zajenkowski, Szymanik, & Garraffa, 2014). Some psychological studies (e.g., Newstead et al., 1987; Ragni, Eichhorn, Bock, Kern-Isberner, & Tse, 2017) have demonstrated that many quantifiers do not have a precise true/false cutoff for the quantity they represent, on a scale from 0 to 100. Even more, the minimum and maximum values of human's subjective valuation responses to individual quantifier can vary a lot (Ragni et al., 2017). This may hint at a fuzzy underlying space of quantifiers among people, in terms of using and interpreting quantifiers (Budescu & Wallsten, 1985).

Generalized quantifiers have been recently employed in studies of syllogisms[1], such as the Probability Heuristics Model (PHM; Oaksford & Chater, 1994). A recent study (Ragni, Singmann, & Steinlein, 2014) has extended three syllogistic reasoning theories (Matching Hypothesis, Mental Model Theory and Preferred Mental Models) to generalized quantifiers. However, only the two quantifiers – "most" and "few" were included. The interpretation of generalized quantifiers plays a role in most reasoning theories, especially regarding the set relationship. For example, in mental model theory, it is the basis for the construction of mental models. Endorsement of invalid conclusion can be resulted when reasoners commit the illicit conversion fallacy of interpreting "All As are Bs" as equivalent to "All Bs are As" and make a mistake in the initial mental model construction. This is indeed a very common error in syllogistic tasks. While for PHM, it is relevant to the selection of the preferred quantifier in the conclusion.

Knowing the factors affecting the underlying representation of generalized quantifiers is essential for cognitive theories for reasoning. One example is that "most As are

[1] Syllogisms are deductive reasoning problems in which one or more conclusions are derived from two premises. The two premises are categorical propositions which are assumed to be true. For example, the conclusion "All As are Cs" can be drawn from the two premises "All As are Bs" and "All Bs are Cs". The abstract terms, A, B and C can be substituted by concrete categorical terms like "apple" and "fruit".

Bs" is "equivalent" to "most Bs are As" if A and B have the same set size. However, if B has a much larger set size or cardinality than A, then the proposition does not hold after the switch (also known as illicit conversion). About 50% of the participants in the experiments of a previous study chose "most" as the conclusion quantifier for a syllogism with "most" as the quantifier for both premises[2] (Chater & Oaksford, 1999). It is interesting that half of the participants considered A, B and C as having the same cardinality while the other half of the participants may not. About 15% of the participants in experiment 1 and 20% of the participants in experiment 2 chose "no valid conclusion" as their responses. It is very possible that these participants may be aware of the fact that the differences in total set sizes of A, B and C can lead to different conclusions for the syllogism.

Factors affecting the variety in the interpretation of quantifiers and underlying space have to be controlled in studies employing these quantifiers. Besides, are there individual differences in the underlying representation of a quantifier? Does the underlying space affect the response diversity? The answers may provide insight for the questions why human participants do not always draw the same or logical conclusions but some particular irrational conclusions are preferred and why more response diversity was found for more difficult reasoning problems (e.g., Khemlani & Johnson-Laird, 2012). Also, would the degree of vagueness/uncertainty of a quantifier cause more individual differences? What are the factors of the response diversity in the interpretation of generalized quantifiers? Will the degree of uncertainty cause a larger processing difficult which can be reflected by a longer processing time? And what are the factors of processing difficulty of generalized quantifiers? This analysis investigates these questions regarding the underlying space and processing time (difficulty) of generalized quantifiers. More precisely, we focus on three levels of tasks: 1. Spontaneous valuation of the quantity or frequency the quantifier represents in the *Subjective Valuation Task*. 2. A *Truth Judgement Task* in which participants were asked to judge if a quantified statement holds true for a picture. 3. Finally, a *syllogistic reasoning task* with generalized quantifiers that participants were asked to reason and derive a conclusion from two premises of quantified statements. Please refer to the Method section for details about the three tasks.

Two measures of response diversity were employed in this study, namely the standard deviation and Shannon entropy. In information theory, Shannon entropy calculates the expected value of the information transmitted in a message (Shannon & Weaver, 1949), as a function of the probability of the occurrence of each possible message. For each quantifier, the entropy in the Truth Judgement Task, was calculated by the aggregated normalized probabilities of the truth responses for each of the pictures/scenarios presented (see Method for the details) by the Shannon equation: $-\sum p_i \log_2 p_i$, where p_i is the probability of a truth response. There will be several "small" probabilities if the response is more diverse. Conversely, if the responses are condensed to a few values, the probabilities of these selected values will be high. The smaller the probability, the larger the entropy value calculated by the equation. And thus, a larger Shannon entropy value indicates more discrepancy in the responses. It was used to measure the response diversity in syllogistic reasoning in a meta-analysis study (Khemlani & Johnson-Laird, 2012). Standard deviations of the responses in the Subjective Valuation Task and Truth Judgement Task were calculated as well to check if the two measures of diversity agree with each other. For answering the question regarding underlying space, Newstead et al. (1987) found that the amount of entities represented by certain quantifiers (e.g., "some") could be affected by the assumed total set size of the experimental scenario. However, does this hold for all generalized quantifiers? Besides the extra-linguistic factor of total set size, several properties of the quantifier itself can also affect the diversity of human responses (interpretation) and response time. They include, among others, quantifier type, monotonicity, and word frequency. We will elaborate two aspects below.

Quantifier Types

There are many different ways of classifying quantifiers (e.g., logical quantifiers versus different types of binary quantifiers; simple versus complex quantifiers). In this study, the quantifiers were classified in the sense of natural language, namely frequency versus quantity quantifiers. Many studies have been conducted for quantity quantifier, while there are only few for frequency ones. For instance, Newstead and Collis (1987) studied the context effect in the interpretation of ten frequency quantifiers. In contrast to some previous findings for quantity quantifiers (Chase, 1969; Newstead & Griggs, 1984; Newstead et al., 1987), no significant set size effect or effect due to the presence of other quantifiers were found. This supported that processing of quantifiers of different types may be different due to their specific properties.

Monotonicity

A generalized quantifier Q_{up} is upward monotone/entailing (or monotone increasing) if and only if for all M and all A, $B \subseteq B' \subseteq M$, $Q_M(A,B)$ implies $Q_M(A,B')$. That means Q_{up} license the inference from subsets to supersets. Similarly, a Q_{down} is downward monotone/entailing if and only if for all M and all A, $B \subseteq B' \subseteq M$, $Q_M(A,B')$ implies $Q_M(A,B)$. Contrastive to Q_{up}, Q_{down} license the inference from supersets to subsets. For example, "Some men are Germans implies some men are Europeans". With the fact that "Germans" is within the set of "Europeans", "some" is an upward monotone quantifier. Similarly, "No men are birds" implies "No men are eagles". With the fact that eagles are birds, "no" is downward monotone. There are non-monotone quantifiers also, e.g., "exactly three". For example, "exactly three men are Germans" does not imply "exactly three men are Europeans" or vice versa. According to

[2] The syllogism mentioned here is the MM4 syllogism in the study: Premise 1 as "Most As are Bs"; Premises 2 as "Most Bs are Cs"; and conclusion as "Most Cs are As".

the definition, many natural language quantifiers are (either upward or downward) monotone, including the three Aristotle quantifiers, "all", "some" and "no". Barwise (1981) suggested that monotone quantifiers are easier to process than non-monotone ones.

Aims of the Study and Research Questions

We aimed to examine human's interpretation of a large number of generalized quantifiers and factors affecting the response diversity and processing time of these quantifiers to facilitate further studies of generalized quantifiers in different fields. As mentioned before, the significant properties have to be controlled in studies involving these quantifiers in order to eliminate some confounding factors. The analyses focus on three factors, namely total set size, quantifier type (quantity quantifier and frequency quantifier), and monotonicity (upward, downward and non-monotone), according to two domains, namely degree of variation in underlying representation space (among interpreters, in terms of response diversity) and processing difficulty (in terms of processing time) of the generalized quantifiers.

Research Question 1: Factors of Response Diversity

Unlike "All", "No" or "Seven" (numerical quantifiers), the amount (or proportion) represented by most generalized quantifiers can be rather fuzzy. Humans do not agree with each other regarding the representation space of individual quantifier. What are the factors affecting the differences in the underlying representation space of quantifiers? In other words, are total set size, quantifier type and monotonicity the factors affecting response diversity? Research question 1 (RQ 1) was examined by analyzing the standard deviation (SD) and Shannon entropy measures in both the Subjective Valuation and Truth Judgement Tasks. We hypothesized that the smaller set size condition, quantity quantifiers and upward monotone quantifiers may exhibit smaller response diversities, i.e., smaller standard deviation and entropy measure values.

Research Question 2: Processing Time

Does greater degree of fuzziness cause a longer processing time (in terms of response time)? Besides, is the monotonicity a factor of processing time as well? Szymanik and Zajenkowski (2013) found a significant interaction effect of monotonicity and the truth value of the quantified statement in a verification task of four quantifiers but failed to find a significant main effect of monotonicity. We extended the study with more quantifiers of different quantifier types. Word frequency was included as a covariant because it has a general effect in word recognition[3]. Quantifiers of higher word frequency are expected to be processed faster. We hypothesized that the entropy measures and word frequency are significant predictors of RT; and the quantifier type may affect the RT as well.

[3] Quantifiers with higher word frequencies are supposed be processed faster due to the availability heuristic or ease of retrieval, having a faster response time in a spontaneous timed task. A significant decrease of the recognition time for words with higher word frequency (e.g., O'Malley & Besner, 2008) is generally found. The word frequency measures were taken from the British National Corpus: http://www.natcorp.ox.ac.uk.

Method

Participants

104 native English speakers (M = 40.8 years; range = 21-75 years; 63 females) participated in the online experiment on Amazon Mechanical Turk. We controlled for one participant from a given computer. They received a nominal fee.

Materials, Design and Procedure

A search for common quantity and frequency quantifiers was performed in Google with the keywords "quantifiers", "frequency quantifiers", "frequency adverbs", "determiners", "how often", "how many", and "how much". 77 generalized quantifiers were selected[4], with 34 frequency quantifiers (frequency adverbs); and 14, 13 and 16 quantity quantifiers which can be used with countable, uncountable and both countable and uncountable nouns (type-both) respectively. Each participant had to perform two tasks:

A Subjective Valuation Task Participants were asked to provide a subjective value of the amount the quantifier represents. They had to move a slidebar to indicate their responses in terms of percentages (from 0% to 100%). Each quantifier was evaluated once.

A Truth Judgement Task Participants had to evaluate the validity of a quantified statement presented above a picture. For the effect of total set size, participants were randomly assigned to *either* the smaller-set group or larger-set group. The number of participants in each group was counterbalanced. For countable and type-both quantifiers, pictures of 10 circles or 100 circles were displayed, with 0 to all of them colored black (see Fig. 1). While for uncountable and type-both quantifiers, pictures of a heap of sand or desert (composed of 10 heaps of sand) were presented with 0 to 100% of the sand or desert colored brown (see Fig. 2).

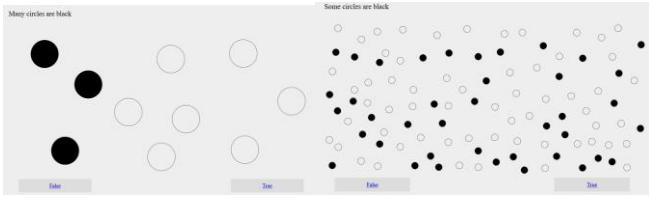

Figure 1: Pictures for countable space in the Truth Judgement Task. Participants received the left (10 circles) *or* right picture (100 circles) and had to evaluate whether a quantified statement like "Some circles are black" (presented at the left hand upper corner) is a true description of the picture or not.

[4] The list can be retrieved from www.cc.uni-freiburg.de/data.

For frequency quantifiers, timelines of a week with a coffee cup icon for 0 to 7 days or a monthly schedule with an icon of football for 0-31 days were presented. Each quantifier was tested 2 times for each participant in two blocks. Pictures displayed were counterbalanced within participants in the sense that if the participant received less than or equal to half positive situation (e.g., 3 out of 10 circles colored black) in the first block, he/she would receive more than or equal to half positive situation in the second block (e.g., 7 out of 10 circles are black) and vice versa. The same manipulation was applied to all the three scenarios (circles, sand/desert, and timeline/calendar). Type-both quantifiers were tested four times for each participant as they were presented twice in both the circle and sand/desert scenarios. The possible picture options were counterbalanced among participants. For countable and type-both quantifiers, the statement was in the form of "*Quantifier* (of the) circles are black" or "These circles are *Quantifier* black" (for "commonly" and "typically"). For the sand/desert situation, the statement was "*Quantifier* sand is colored brown". The statement was in the form of "Tim *Quantifier* drinks coffee" or "Tim drinks coffee *Quantifier*" for the weekly timeline scenario; and "Tim *Quantifier* plays soccer" or "Tim plays soccer *Quantifier*" for the monthly calendar scenario. Participants were asked to judge as accurately and quickly as possible whether the statement was a truth description of the picture. Participants always performed the Subjective Valuation Task first.

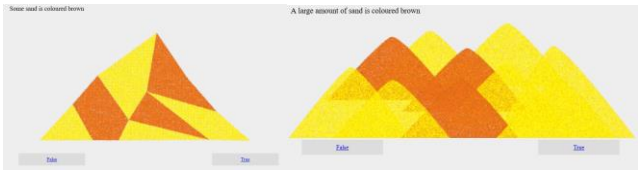

Figure 2: Pictures for an uncountable space in the Truth Judgement Task. Participants had to judge whether a statement like "Most sand is colored brown" is a true description of the picture or not.

Results

The Underlying Representation Space

For the first research question, the diversity in the responses was evaluated by the Shannon entropy and standard deviation measures of the responses in both the Truth Judgement Task and Subjective Valuation Task, as the indices. The standard deviation and entropy measures of the responses were calculated according to the two different set size conditions in the Truth Judgement Task (SD1, Entropy1; and SD2, Entropy2). SD1 and Entropy1 are the standard deviation and entropy measure of the smaller set size pictures (10 circles, 1 heap of sand and weekly timeline). SD2 and Entropy2 are the standard deviation and entropy measure of the larger set size condition (100 circles, desert and monthly schedule). The Spearman's rank correlations between the three SDs and entropy measures were tested both within and across the two tasks. Significant correlations were found except for SD2 with the entropy in the Subjective Valuation Task and Entropy1, see table 1 for the results. For the effect of set size on the response diversity, significant differences were found between both SD1 and SD2; and Entropy1 and Entropy2, $t(76) = -6.142$, $p < .001$ and $t(76) = -7.268$, $p < .001$, respectively, with SD2 and Entropy2 being significantly larger. The two measures (SD and entropy) provided similar results, as the entropy measures were more reliable indices for the response diversity (according to the positive correlations across tasks), we used the entropy measures for the following analyses for the sake of simplicity.

Regarding the property of monotonicity, the quantifiers were classified into 28 upward monotone, 23 downward monotone, 11 monotone and 12 non-monotone quantifiers. The effects of the three quantifier properties on the two entropy measures (for underlying space) in the Truth Judgement Task were then tested. The 2 (quantifier type: frequency and quantity) x 4 monotonicity (monotonicity: upward, downward, monotone and non-monotone) MANOVA, with word frequency as a covariate, for the two entropy measures showed that the quantifier type had a significant multivariate effect on the two entropy measures, $F(2, 64) = 179.820$, $p < .001$, Wilk's $\lambda = .151$, $\eta_p^2 = .849$; as well as monotonicity and word frequency, $F(6, 128) = 5.568$, $p < .001$, Wilk's $\lambda = .629$, $\eta_p^2 = .207$ and $F(2, 64) = 10.130$, $p < .001$, Wilk's $\lambda = .760$, $\eta_p^2 = .240$, respectively. The interaction effect of quantifier type and monotonicity was not significant.

The following post-hoc tests were performed according to quantifier type (frequency versus quantity) and monotonicity (upward and downward). The t-tests showed that the two entropy measures (Entropy1 and Entropy2) were reliably different for frequency quantifiers, $t(33) = -23.426$, $p < .001$, but not for quantity quantifiers. Regarding the monotonicity, the two entropy measures were reliably different for upward and downward monotone quantifiers, Entropy1: $t(49) = 2.328$, $p = .024$; Entropy2: $t(49) = 3.198$, $p = .002$.

Differences between the response diversity indices for the first half and second half of the tasks were also examined. Regarding Entropy1 and Entropy2, both t-tests were not significant, Entropy1: $t(76) = 1.236$, $p = .220$, Entropy2: $t(76) = .455$, $p = .650$. For the Subjective Valuation Task, there was no difference between the first and second half of the task neither, $t(73) = -.067$, $p = .946$.

Processing Time

We filtered out the response times which exceed average RT +/- 2 SD according to individual participant. Firstly, a stepwise regression was performed to test if the word frequency and the three entropy measures significantly predicted the response time. The results of the regression analysis showed that the two entropy measures in the Truth Judgement Task explained 38.6% of the response time (adjusted $R^2 = .386$, $F(2, 74) = 24.928$, $p < .001$, Entropy1: $\beta = .975$, $p < .001$, Entropy2: $\beta = -.603$, $p < .001$. As the response times of the two quantifier types were significantly different,

Table 1: Results of Spearman's rank correlation of the standard deviation (SD) and entropy measures of the responses in the Subjective Valuation Task (SD and entropy) and Truth Judgement Task (SD1 and SD2; and Entropy1 and Entropy2).

		Valuation	Judgement			
		SD	Entropy1	Entropy2	SD1	SD2
Valuation	entropy	-.359**	.566**	.425**	.450**	0.188
	SD		-.257*	-.326**	-.485**	-.444**
Judgement	Entropy1			.503**	.483**	-0.086
	Entropy2				.557**	.599**
	SD1					.555**

*Correlation is significant at the .05 level (2-tailed).
**Correlation is significant at the .01 level (2-tailed).

$t(75) = 7.188$, $p < .001$, the regression was repeated according to the two quantifier types. For frequency quantifiers, word frequency was the only significant predictor, adjusted $R^2 = .157$, $F(1, 32) = 7.128$, $p = .012$; $\beta = -.427$, $p = .012$. While for quantity quantifiers, entropy in the Subjective Valuation Task was the only significant predictor of the response time, adjusted $R^2 = .136$, $F(1, 41) = 7.609$, $p = .009$; $\beta = .396$, $p = .009$. The effect of monotonicity on response time was not significant. Do generalized quantifiers affect the response diversity in syllogistic reasoning? Syllogisms are chosen as quantifiers are the essence of syllogistic reasoning and so their effect may be most visible.

Entropy in Reasoning with Generalized Quantifiers: Additional Empirical Support

We reanalyzed the data from Ragni et al. (2014) with the entropy measure for response diversity. Twenty-five native English speakers participated in the online experiment on Amazon Mechanical Turk. Each participant had to solve 40 syllogistic problems with at least one of "most" and "few" being the quantifier of one of the two premises. Participants had to choose the conclusion quantifier of the syllogism among the four classical Aristotle quantifiers and the two generalized quantifiers "most" and "few" (i.e., all, no, some, some…not, most and few), to the question "what follows?" after reading the two premises. The conclusion direction presented (a-c or c-a) was counterbalanced. 20 problems were tested for each conclusion direction. "Most" and "few" appeared in the first premise respectively in 6 of the syllogisms, with the second premise being one of the six quantifiers (6 x 2 = 12 problems). For the 8 remaining syllogisms, "most" and "few" appeared in the second premise, with the first premise being one of the four Aristotle quantifiers.

The entropy measure of the responses for each syllogism was calculated and an ANOVA and a t-test were performed. The 2 (conclusion direction: a-c vs. c-a) x 2 (position of the generalized quantifier: first premise vs. second premise) ANOVA showed a significant main factor of the position, $F(2, 39) = 4.738$, $p = .015$, $\eta_p^2 = .218$, but both conclusion direction and the interaction effect were not significant. Post-hoc analysis showed that syllogisms with generalized quantifier in the first premise had a significantly higher entropy, $t(30) = 2.174$, $p = .038$ (2-tailed). The number of generalized quantifier affects the entropy as well. If both premises contained generalized quantifiers, the entropy was marginally smaller, Independent Samples Test: $t(38) = 1.957$, $p = .058$ (2-tailed)[5]. The marginal result might due to the fact that only 8 syllogisms have two generalized quantifiers but 32 problems have only one generalized quantifiers.

General Discussion

While extensive research in psychology of reasoning and logic has dealt with the four classical quantifiers ("all", "some", "some…not", and "none"), few cognitive reasoning theories for syllogisms have been extended to generalized quantifiers – and often to "most" and "few" only. Different quantifiers possess different specific properties which affect their interpretation (especially in terms of interpretation diversity) in daily language. For example, for universal quantifiers like "All" and "No", most participants would select 100% and 0%, respectively, in the Subjective Valuation Task, with few selecting values within 95% to 99% and 0% to 5%, respectively. In contrast, the more "fuzzy" generalized quantifiers elicit a greater diversity in the responses. For example, for "some", we got 6 responses for 20% and 35%, 5 responses for 25%, and 10 responses for 45% (among 104 responses). In total, 47 different percentages (out of 101 possible choices) were selected as the responses in the Subjective Valuation Task. It seems that the right tool to analyze the interpretation diversity is missing. We argue that Shannon entropy, which was developed for communication, is an excellent method which can be employed to measure the response diversity of generalized quantifiers.

Using Shannon's entropy to measure response diversity was introduced in this study as it is a binary-based element which fits the dichotomous experimental design of the Truth Judgement Task. It shows reliable correlated results with the classical standard deviation measure within and across tasks. Shannon entropy seems a better measure for response diversity across tasks than the SD. Our results show that the total set size, quantifier types (frequency versus quantity) and monotonicity can affect the interpretation diversity of a quantifier; while the interpretation diversity (in terms of

[5] We performed the Levene's test for equality of variances and the results were not significant, i.e. equal variance can be assumed.

entropy measures) can in turn affect the response time. In accordance with the findings of Szymanik and Zajenkowski (2013), we did not find a reliable effect of the monotonicity of quantifiers on RT. One can speculate that the difference in processing difficulty applies to cardinal quantifiers only.

The smaller set size condition has a smaller entropy value as hypothesized, in contrast to the frequency quantifiers and downward monotone quantifiers. Further studies are required to explain this finding. Quantity quantifiers have a slower RT in general and this might be affected by the larger discrepancy in the underlying representation space, which hints a fuzzier underlying representation of quantity quantifiers *among* participants. However, our results suggest that quantifiers possess a rather static underlying mental representation space *within* participants, not changing within or across tasks, as there is no difference for the response diversity measures between the first and second half of the tasks.

Despite our finding of total set size being a factor in the interpretation of generalized quantifiers, it is still possible for human to interpret quantifiers without the knowledge of total set size (Van Tiel & Geurts, 2013). But we can speculate that participants usually represent the underlying set by a default mental model for the respective quantifiers.

Our study shows that total set size, quantifier type, and monotonicity (and word frequency) are all contributing to the possible diversity in the use or reasoning of generalized quantifiers. Based on these factors, natural extensions of theories which already assume models of different sizes and are analogous representations of the state of affairs (like the mental model theory) to incorporate the proposed results is possible. Extension to generalized quantifiers is increasingly important for cognitive reasoning theories to avoid a self-centered focus, which renders them ultimately useless for explaining or predicting complex everyday communication. Large-scale studies of these generalized quantifiers in reasoning tasks can test if the diversity in the interpretation of these quantifiers is the factor of the response diversity in reasoning tasks. It is possible that differences in the interpretation of the quantifier contribute to the deviation from logical responses, other than reasoning/heuristic processes. Controlling the above significant factors is important for studies involving quantifiers, to avoid hidden experimental confounds. Also, for theories with predictions on response time, it is possible that interpretation diversity is a significant factor. Further studies on this hypothesis are necessary.

Acknowledgement

This work is supported by DFG-Projects under grant numbers RA 1934/2-1, RA 1934/3-1, and RA 1934/4-1.

References

Austin, J. L., Strawson, P. F., Grice, H. P., Chomsky, N., Katz, J. J., Goodman, N., et al. (1971). *The philosophy of language (Vol. 39).* London: Oxford University Press.

Barwise, J., & Cooper, R. (1981). Generalized Quantifiers and Natural Language. *Linguistics and Philosophy, 4*(2), 159-219.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental psychology, 58*(5), 412-424.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36*(3), 391-405.

Chase, C. I. (1969). Often is where you find it. *American Psychologist, 24*(11), 1043.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive psychology, 38*(2), 191-258.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin, 138*(3), 427-457.

Lindström, P. (1966). First order predicate logic with generalized quantifiers. *Theoria, 32*(3), 186-195.

Moxey, L. M., & Sanford, A. J. (1993). Prior expectations and the interpretation of natural language. *European Journal of Cognitive Psychology, 5*(1), 73-91.

Newstead, S. E., Pollard, P., & Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Applied ergonomics, 18*(3), 178-182.

Newstead, S., & Collis, J. M. (1987). Context and the interpretation of quantifiers of frequency. *Ergonomics, 30*(10), 1447-1462.

O'Malley, S., & Besner, D. (2008). Reading aloud: Qualitative differences in the relation between stimulus quality and word frequency as a function of context. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1400-1411.

Ragni, M., Eichhorn, C., Bock, T., Kern-Isberner, G., & Tse, A. P. P. (2017). Formal Nonmonotonic Theories and Properties of Human Defeasible Reasoning. *Minds & Machines, 27*(1), 37-77.

Ragni, M., Singmann, H., & Steinlein, E. M. (2014). Theory Comparison for Generalized Quantifiers. *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1330-1335). Austin, TX: Cognitive Science Society.

Segui, J., Mehler, J., Frauenfelder, U., & Morton, J. (1982). The word frequency effect and lexical access. *Neuropsychologia, 20*(6), 615-627.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana: University of Illinois Press.

Szymanik, J., & Zajenkowski, M. (2013). Monotonicity has only a relative effect on the complexity of quantifier verification. *Proceedings of the 19th Amsterdam Colloquium*, (pp. 219-225). Amsterdam.

Van Tiel, B., & Geurts, B. (2013). Truth and typicality in the interpretation of quantifiers. *Proceedings of Sinn und Bedeutung 18*, (pp. 451-468). Basque Country.

Zajenkowski, M., Szymanik, J., & Garraffa, M. (2014). Working memory mechanism in proportional quantifier verification. *Journal of psycholinguistic research, 43*(6), 839-853.