Generalized Quantifier Distribution and Semantic Complexity

Camilo Thorne
KRDB Research Centre
Bolzano, Italy
cthorne@inf.unibz.it

Jakub Szymanik
Institute for Logic, Language and Computation
Amsterdam, The Netherlands
J.K.Szymanik@uva.nl

Abstract

We study in this paper the correlation between the semantic (data) complexity of first order, proportional, aggregate and Ramsey generalized quantification and their distribution in corpora. We show that, as expected by the theory, such distribution is significantly skewed towards tractable as opposed to intractable quantifiers. We also study whether such relationship can be described by a power law.

1 Introduction

Quantification is an essential feature of natural languages. It is used to specify the (vague) number or quantity of objects satisfying a certain property. Quantifier expressions are built from *noun phrases* (whether definite or indefinite, names or pronouns) and *determiners* resulting in expressions such as “a subject”, “more than half of the men”, “the queen of England”, “John”, “some”, “five” or “every”.

More recently, interest has arised regarding *semantic complexity*, viz., the complexity of reasoning with (and understanding) natural language quantifiers —a problem of interest for both cognitive science and computational linguistics. One model that has been proposed to study natural language semantic complexity is to consider the computational properties that arise from formal semantic analysis, see e.g., [11, 2, 14].

Generalized Quantifier Theory (see, e.g., [9]) makes it also possible to distinguish *tractable* from *intractable* quantifiers, see, e.g., [3, 15]. Following some work in cognitive science, e.g., [5, 12] one would expect that speakers (due to their restricted cognitive resources) are naturally biased towards low complexity expressions (tractable or lower) [15, 8].

Related work by the first author in [17] shows that, when one considers the satisfiability problem of English sentences (specifically, of its fragments possessing a first-order semantics, see, e.g., [10]), then tractable combinations of first-order English constructs occur exponentially more frequently than intractable (or undecidable) ones.

This paper extends such work to the domain of quantifiers and contributes to the semantic complexity debate by focusing on the verification problem and on how the ensuing computational properties affect their distribution in corpora. Specifically, we show for a selected set of corpora that quantifier distribution is skewed towards tractable quantifiers.

2 Semantic Complexity of Generalized Quantifiers

Generalized quantifier theory models the meaning of determiners with first order (FO) and higher order (HO) logics augmented by generalized quantifiers. We can think about generalized quantifiers as relations between subsets of a given universe. For instance, in a given model $\mathcal{I} = (\mathbb{D}_{\mathcal{I}}, \cdot^{\mathcal{I}})$ the statement “most As are B” says that $|A^{\mathcal{I}} \cap B^{\mathcal{I}}| > |A^{\mathcal{I}} \setminus B^{\mathcal{I}}|$, where $A^{\mathcal{I}}, B^{\mathcal{I}} \subseteq \mathbb{D}_{\mathcal{I}}$. Going a step further we can take a generalized quantifier Q to be a functional relation associating with each model $\mathcal{I}$ a relation between relations on its universe, $\mathbb{D}_{\mathcal{I}}$. This is actually equivalent to their standard definition from model theory, where generalized quantifiers are simply classes of models:

Definition 2.1. Let $t = (n_1, \ldots, n_k)$ be a k-tuple of positive integers. A *generalized quantifier* of type t is a class Q of models of a vocabulary $\tau_t = \{R_1, \ldots, R_k\}$, such that R_i is n_i-ary for $1 \leq i \leq k$, and Q is closed under isomorphisms, i.e. if $\mathcal{I} \in Q$ and $\mathcal{I}'$ is isomorphic to $\mathcal{I}$, then $\mathcal{I}' \in Q$.

Each generalized quantifier Q of type t over vocabulary $\tau_t = \{R_1, \ldots, R_k\}$ gives rise to a so-called (FO or HO) *query* $Q(R_1, \ldots, R_k)$, viz., a (FO or HO) formula (possibly grounded or closed) over τ_t, such that $\mathcal{I} \in Q$ iff $\mathcal{I} \models Q(R_1, \ldots, R_k)$, for all interpretations over τ_t (viz., $Q(R_1, \ldots, R_k)$ *expresses* Q). The complexity of a quantifier can be determined by considering the logical verification problem that arises:

Definition 2.2. The *model checking* problem is the following decision problem. **Input:** A finite model $\mathcal{I}$ over $\tau_t = \{R_1, \ldots, R_k\}$, and a generalized quantifier Q of type t. **Question:** Does $\mathcal{I} \models Q(R_1, \ldots, R_k)$?

When considering (finite) model checking we are interested in its complexity w.r.t. the size of the model, that is, in *data complexity* [6]. The data complexity of model checking induces a partition into *tractable* and *intractable* generalized quantifiers. Respectively: quantifiers, for which model-checking is *at most* **P**, and quantifiers that are exponential, viz., for which model-checking is *at least* **NP**-hard.

2.1 Tractable Quantifiers.

First-order. First-order quantifiers Q of type t over $\tau_t = \{R_1, \ldots, R_n\}$ are quantifiers which give rise to FO queries $Q(R_1, \ldots, R_n)$. They are the best known and most thoroughly studied, and also those with the lowest complexity: model checking in FO (with identity) is in $\mathbf{AC}^0$. See Table 1.

Table 1: Base tractable FO, proportional, and aggregate quantifiers studied in this paper. The other quantifiers mentioned in Section 3 can be defined from these by complementation and duality.

Quantifier	Model Class	D.C.	Example
some	$\{\mathcal{I} \mid (A^{\mathcal{I}}, B^{\mathcal{I}}) \subseteq \mathbb{D}_{\mathcal{I}} \times \mathbb{D}_{\mathcal{I}}$ and $A^{\mathcal{I}} \cap B^{\mathcal{I}} \neq \emptyset\}$	**AC**^{0}	some men are happy
at least k	$\{\mathcal{I} \mid (A^{\mathcal{I}}, B^{\mathcal{I}}) \subseteq \mathbb{D}_{\mathcal{I}} \times \mathbb{D}_{\mathcal{I}}$ and $\lvert A^{\mathcal{I}} \cap B^{\mathcal{I}} \rvert \geq k\}$	**AC**^{0}	at least 5 men are happy
most	$\{\mathcal{I} \mid (A^{\mathcal{I}}, B^{\mathcal{I}}) \subseteq \mathbb{D}_{\mathcal{I}} \times \mathbb{D}_{\mathcal{I}}$ and $\lvert A^{\mathcal{I}} \cap B^{\mathcal{I}} \rvert > \lvert A^{\mathcal{I}} \setminus B^{\mathcal{I}} \rvert\}$	**L**	most trains are safe
$> p/k$	$\{\mathcal{I} \mid (A^{\mathcal{I}}, B^{\mathcal{I}}) \subseteq \mathbb{D}_{\mathcal{I}} \times \mathbb{D}_{\mathcal{I}}$ and $\lvert A^{\mathcal{I}} \cap B^{\mathcal{I}} \rvert \geq p \cdot (\lvert A \rvert / k)\}$	**L**	more than 2/3 of planets are lifeless
total α of	$\{\mathcal{I} \mid (A^{\mathcal{I}}, B^{\mathcal{I}}) \subseteq \mathbb{D}_{\mathcal{I}} \times \mathbb{R}$ and $\mathbf{sum}(\mu_\alpha(A^{\mathcal{I}})) \in B^{\mathcal{I}}\}$	**L**	the total surface of shopping centers is hard to measure
number of	$\{\mathcal{I} \mid (A^{\mathcal{I}}, B^{\mathcal{I}}) \subseteq \mathbb{D}_{\mathcal{I}} \times \mathbb{N}$ and $\lvert A^{\mathcal{I}} \rvert \in B^{\mathcal{I}}\}$	**L**	the number of weeks in a year is even
α-est	$\{\mathcal{I} \mid (A^{\mathcal{I}}, B^{\mathcal{I}}) \subseteq \mathbb{D}_{\mathcal{I}} \times (\mathbb{R} \cup \mathbb{D}_{\mathcal{I}})$ and $\mu_\alpha(\{d\}) = \mathbf{max}(\mu_\alpha(A^{\mathcal{I}})) \in B^{\mathcal{I}}\}$	**L**	the highest mountain in Peru is the Huascaran
average α of	$\{\mathcal{I} \mid (A^{\mathcal{I}}, B^{\mathcal{I}}) \subseteq \mathbb{D}_{\mathcal{I}} \times \mathbb{R}$ and $\mathbf{avg}(\mu_\alpha(A^{\mathcal{I}})) \in B^{\mathcal{I}}\}$	**L**	the average height of mountains in Peru is 5,000 metres

Proportional. More interesting are *proportional* quantifiers. The former, studied extensively in the literature, are *most* ("most men") and $> p/k$ ("more than one third of men"), and their relatives.

Aggregate. Less known but equally interesting are *aggregate* quantifiers such as: *total α of* or, simply, *sum* ("the total surface of"); *number of* or, simply, *count* ("the number of men"); *average α of* or, simply, *avg* ("the average weight"); and *α-est* or, simply, *min* or *max* ("the hottest day", "the hottest summer"). These quantifiers involve computing a (integer, rational or real) number over the relations of the quantifiers via a (second-order) *aggregate function* such as **sum**[1]; α stands for a *metric* (denoted by a *metric adjective*), viz., a function $\mu_\alpha(\cdot)$ mapping relations A to sets (so-called "aggregates") $\mu_\alpha(A)$ of numbers relatively to attribute α over which aggregation is performed. For instance, once such metric might be, e.g., height or weight. See Table 1.

As proportional quantifiers can be verified using the same algorithm, albeit with some minor variations, it follows that both are in **L** [16][2].

2.2 Intractable Quantifiers

Intractable quantifiers can be derived from tractable ones via various model-theoretic operations, usually marking the transition from monadic to polyadic quantification. Three such operations have been defined in the literature: branching and Ramseyfication [15] as well as various form of type-lifting from distributive to collective quantification [7]. In the present paper we focus on Ramseyfication, that turns a monadic quantifier of type $(1,1)$ into a polyadic quantifier of type $(1,2)$:

Definition 2.3 (Ramseyfication)**.** Let Q be a quantifier of type $(1,1)$. The *Ramseyfication* of Q is the following quantifier R_Q of type $(1,2)$: $R_Q = \{\mathcal{I} \mid (A^{\mathcal{I}}, R^{\mathcal{I}}) \subseteq \mathbb{D}_{\mathcal{I}} \times (\mathbb{D}_{\mathcal{I}} \times \mathbb{D}_{\mathcal{I}})$ s.t. exists $X \subseteq A^{\mathcal{I}}$ s.t. $Q(A^{\mathcal{I}}, X)$ and for all $x, y \in X, (x,y) \in R^{\mathcal{I}}\}$.

Ramseyfication can be conveyed by some English sentences with the reciprocal expression "each other" under a default (strong) interpretation [4, 15, 13]. It intuitively states that the models of the resulting Ramseyfied quantifiers are graphs with connected components. Intractability arises when its application gives rise to a so-called "clique" quantifier: if Q is the quantifier *at leastk* or $> p/k$, then R_Q is **NP**-complete [15]. The downward monotone counterparts, like *at most k* ("at most five men"), $< k$ ("less than 5 men") or $< p/k$ ("fewer than one third of men"), are also intractable.

Notice however that Ramseyfication does not always produce intractable quantifiers. When applied to an "Aristotelian" quantifier (i.e., *some* or *all*) or the "bounded" proportional quantifier *most*, it yields **P** quantifiers [15].

3 Pattern-based Corpus Analysis

In this section we summarize our analysis regarding the occurence of generalized quantifiers (FO, proportional and reciprocal, aggregate) in English question and sentence corpora. We identified such quantifiers indirectly, via part-of-speech (POS) patterns that approximate their surface forms. We considered Penn Treebank/Brown corpus POSs[3]. We considered: **(i)** The base quantifiers from Table 1, plus the FO *all* ("all men", "everybody", "everything", "everyone"), $< k$ and $> k$ ("more than 5 men"), *exacly k* ("5 men", "exactly 3 bridges"); the proportional $> k/100$ ("more than 10% of") and $< p/k$; and the reciprocal *recip* Ramseyfier "each other". **(ii)** The tractable *most+recip*, *all+recip*, *some+all* Ramsey quantifiers, and the intractable $<$ *k+recip*, $>$ *k+recip* and $> p/k$*+recip* Ramsey quantifiers.

Due to reasons of space, we will not specify completely our patterns here, but instead illustrate them with examples. To the *some* the pattern "*.* (someone/pn|somebody/pn |something/pn|some/dti|a/at|many/ap) .**" (i.e., a simple regular expression) was associated. To the reciprocal *recip* (i.e., "each other"), the pattern "*.* each/dt other/ap .**" was associated. Finally, for quantifiers such as *some+recip*, we checked for sentences that match *at the same time* the two regular expressions of their constituent quantifiers.

[1] Technically speaking, in order to capture aggregate quantifiers Definition 2.1 would need to be extended to models equipped with some measure function and numerical sort.

[2] Note that **AC**^{0} $\subseteq$ **L** $\subseteq$ **P**; **AC**^{0} denotes the class of problems solvable via circuits of constant depth, and **L** the class of problems solvable in logarithmic space.

[3] For the POS tagging a a 3-gram tagger, with 2-gram and unigram backoffs, trained over the (POS annotated) Brown corpus, and with 80% accuracy.

Table 2: Corpora used in this study.

Corpus	Size	Domain	Type
Brown	19,741 sent.	Open (news)	Decl.
Geoquery	364 ques.	Geographical	Int.
Clinical ques.	12,189 ques.	Clinical	Int.
TREC 2008	436 ques.	Open	Int.

Table 3: Ramseyfied quantifier (raw) frequencies.

Corpus	$> k$+ *recip*	$> p/k$+ *recip*	*most*+ *recip*	*some*+ *recip*	*all*+ *recip*	$< k$+ *recip*
Brown	1	1	2	2	2	16
TREC 2008	0	0	0	0	0	0
Geoquery	0	0	0	0	0	0
Clinical ques.	0	0	0	0	0	0
total	1	1	2	2	2	16

Using such patterns we observed the frequency of **(i)** generalized quantifiers, and **(ii)** Ramseyfied and non-Ramseyfied counting and proportional quantifiers, to see whether such distribution is skewed towards low complexity quantifiers.

Corpora. To obtain a representative sample, we considered corpora of multiple domains and with sentences of arbitrary type (declarative and interrogative). We considered: **(i)** a subset (A: press articles) of the Brown corpus[4]; **(ii)** a subset (Geoquery880) of the Geoquery corpus[5]; **(iii)** a corpus of clinical questions[6]; and **(iv)** a sample from the TREC 2008 corpus[7]. Table 2 summarizes their main features.

Power Law Behavior, Skewness and χ^2 Tests. We seeked to infer a power law or Zipfean relation between *quantifier frequency* $fr(Q)$ and *quantifier complexity rank* $rk(Q)$, viz.,

$$fr(Q) = a/rk(Q)^{-b}, \qquad (1)$$

where rank refers to how "easy" Q is (i.e., how low its data complexity is). To approximate the parameters a and b it is customary to run a least squares linear regression, since (1) is equivalent to a linear model on the log-log scale. The R^2 coefficient measures how well the observations fit the inferred power law equation.

Power laws are exponential, non-normal and skewed distributions where the topmost (w.r.t. rank) 20% outcomes of a variable concentrate 80% of the probability mass or frequency. They are widespread in natural language data [1].

To validate our models, we run a χ^2 test (at $p = 0.01$ significance) w.r.t. a uniform null hypothesis and measured the overall skewness of the distribution.

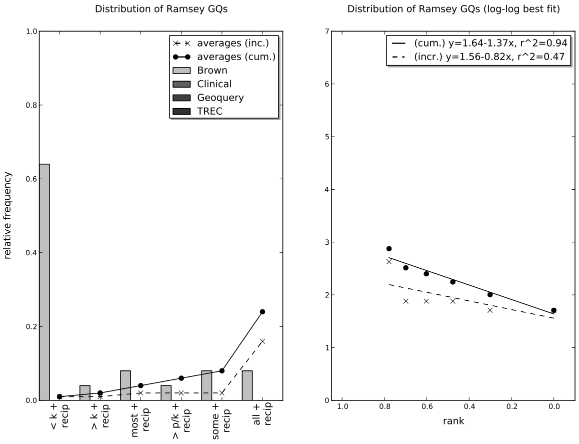

Figure 1: Ramsey Quantifier distribution and log-log power law regression.

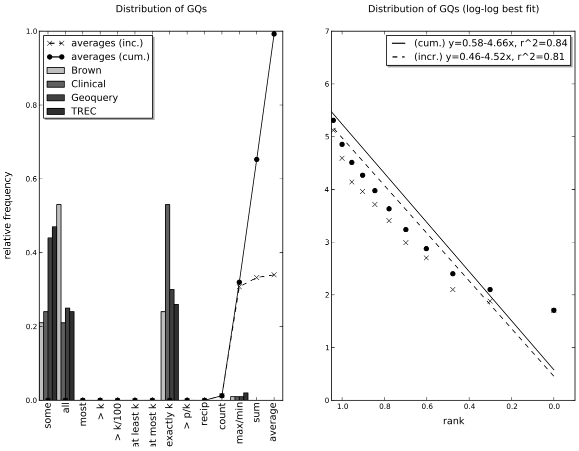

Figure 2: Quantifier distribution and log-log power law regression.

4 Results and Interpretation

The distributions observed are summarized by Figures 1 and 2. Tables 3 and 4 summarize, resp., the contingency raw frequency tables from which the Figures were generated. The reader will find on the left of each figure the relative average and cumulative frequency plots for the quantifiers considered, and to the right the plots of the log-log regressions. In Table 5 we spell out the power law/Zipfean relations inferred, plus the model validation results.

As expected by both the theory and the similar results regarding language fragments, **AC**^{0} quantifiers occur more frequently than **L** (i.e.proportional and aggregate), and their distribution seems to follow a power law (with a high — $R^2 = 0.81$— correlation coefficient) as Figure 2 and Table 5 show. Table 5 shows that this bias is statistically significant: their distribution significantly differs from uniform or random distributions (the null hypothesis rejected by the

Table 5: Summary of test results.

Model/Test	Recip. GQs	GQs
P. law $fr(Q)$	$36.00/rk(Q)^{-0.82}$	$2.88/rk(Q)^{-4.52}$
Skewness	1.76	1.98
χ^2 value	530.81	$183, 815, 415, 173.11$
p value, df.	1.78, 5	0.0, 13
R^2 coeff.	0.46	0.81

[4] http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml

[5] http://www.cs.utexas.edu/users/ml/nldata/geoquery.html

[6] http://clinques.nlm.nih.gov

[7] http://trec.nist.gov

Table 4: Observed quantifier (raw) frequencies.

Corpus	*at least k*	$< k$	*most*	$> k$	$> p/k$	*recip*	$> k\%$	*sum*	*count*	*avg*	*max/min*	*all*	*exactly k*	*some*
Brown	192	4	1532	540	38	101	2	1	354	17	4368	202587	90811	81693
TREC 2008	0	0	0	0	0	0	0	0	0	0	13	192	490	222
Geoquery	2	0	0	0	0	0	0	0	1	0	18	380	447	660
Clinical ques.	12	0	28	12	0	0	0	0	9	2	889	10712	11629	20780
total	206	4	1560	552	38	101	2	1	364	19	5288	213871	103377	103355

test), since $p < 0.01$. Their distribution shows also a high measure of skewness. Furthermore, a comparison of Table 3 with Table 4 shows that such tractable quantifiers occur exponentially more often that intractable (viz., intractable Ramseyfied) quantifiers. This fact is further substantiated by the comparatively very low (raw and relative) frequency of *recip* (see again Figure 2 and Table 4).

Regarding Ramseyfied quantifiers themselves, viz., tractable vs. intractable Ramseyfied quantifiers, the results were less conclusive. The distribution shows a bias towards intractable Ramseyfications, but does not exhibit (see Figure 1) a conclusive power law/Zipfean relation; the mean relative frequency regression yielded a rather low ($R^2 = 0.46$) correlation coefficient. Furthermore, the bias was not significant enough to reject the null hypothesis (i.e., a uniform distribution), even if skewness remains high; see Table 5. These inconclusive results might be due to the general "sparseness" of such quantifiers: as Table 4 shows, Ramseyfications are in general rare in natural language data. It is difficult to gather sufficient data to observe clear statistical trends among them, although the data we collected strongly suggests that they are (altogether) less frequent than non-Ramseyfied $\mathbf{AC}^0$ and $\mathbf{L}$ FO, proportional and aggregate quantifiers.

These results seem however to show altogether—as suggested by [11, 8]—that: even though an everyday fragment of natural language allows some intractable constructions, it consists mostly of tractable expressions. Moreover, the computationally easier the expression the more often it is used in everyday communication.

5 Conclusions

We have studied the distribution of FO, proportional, aggregate and Ramsey generalized in corpora, to observe if their distribution is correlated to their semantic (data) complexity. In particular, to observe if, as expected by the theory of and related results in semantic complexity, such distribution is significantly skewed towards tractable as opposed to intractable quantifiers. Our results suggest that this is indeed the case. Furthermore, the distribution of $\mathbf{AC}^0$ vs. $\mathbf{L}$ quantifiers seems to follow a power law. We obtained less conclusive results regarding tractable and intractable Ramsey quantifiers due to their overall "sparseness".

References

[1] Marco Baroni. Distributions in text. In Mouton de Gruyter, editor, *Corpus linguistics: An International Handbook*, volume 2, pages 803–821. 2009.

[2] J. Van Benthem. Towards a computational semantics. In Peter Gärdenfors, editor, *Generalized Quantifiers*, pages 31–71. Reidel Publishing Company, 1987.

[3] A. Blass and Y. Gurevich. Henkin quantifiers and complete problems. *Annals of Pure and Applied Logic*, 32:1–16, 1986.

[4] M. Dalrymple, M. Kanazawa, Y. Kim, S. Mchombo, and S. Peters. Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy*, 21:159–210, 1998.

[5] Marcello Frixione. Tractable competence. *Minds and Machines*, 11(3):379–397, 2001.

[6] Neil Immerman. *Descriptive Complexity*. Texts in Computer Science. Springer, New York, NY, 1998.

[7] Juha Kontinen and Jakub Szymanik. A remark on collective quantification. *Journal of Logic, Language and Information*, 17(2):131–140, 2008.

[8] Marcin Mostowski and Jakub Szymanik. Semantic bounds for everyday language. *Semiotica*, 188(1-4):363–372, 2012.

[9] S. Peters and D. Westerståhl. *Quantifiers in Language and Logic*. Clarendon Press, Oxford, 2006.

[10] Ian Pratt-Hartmann. Computational complexity in natural language. In Alex Clark, Chris Fox, and Shalom Lappin, editors, *Computational Linguistics and Natural Language Processing Handbook*. Blackwell, 2008.

[11] Eric S. Ristad. *The Language Complexity Game*. Artificial Intelligence. The MIT Press, March 1993.

[12] Iris Van Rooij. The tractable cognition thesis. *Cognitive Science: A Multidisciplinary Journal*, 32(6):939–984, 2008.

[13] F. Schlotterbeck and O. Bott. Easy solutions for a hard problem? the computational complexity of reciprocals with quantificational antecedents. pages 60–72. CEUR Workshop Proceedings, 2012.

[14] J. Szymanik. *Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language*. PhD thesis, University of Amsterdam, Amsterdam, 2009.

[15] Jakub Szymanik. Computational complexity of polyadic lifts of generalized quantifiers in natural language. *Linguistics and Philosophy*, 33(3):215–250, 2010.

[16] Camilo Thorne. *Query Answering over Ontologies Using Controlled Natural Languages*. PhD thesis, Faculty of Computer Science, 2010.

[17] Camilo Thorne. Studying the distribution of fragments of english using deep semantic annotation. In *Proceedings of the ISA8 workshop*, 2012.