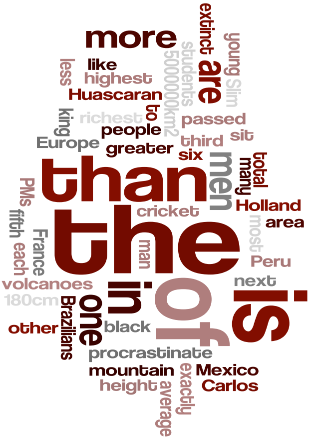
Quantifier Distribution and Semantic Complexity

Camilo Thorne[1] Jakub Szymanik[2]

[1] KRDB Research Centre for Knowledge and Data
cthorne@inf.unibz.it
http://www.inf.unibz.it/~cathorne

[2] Institute for Logic, Language and Computation
jakub.szymanik@gmail.com
http://www.jakubszymanik.com/

TbiLLC2013, Sep 26, Tbilisi



- Words and structures in English occur following some general laws
- A distribution describes how often they occur/probable they are
- E. Zipf showed that in many cases such distributions correspond to power laws

Hypothesis

Quantifiers are power-law distributed w.r.t. semantic complexity

Motivation (ctd.) I

the total area of Europe is greater than 5,000,000 km2

the highest mountain in Peru is the Huascaran

the average height of men in France is 180 cm

less than one fifth of Brazilians like cricket

the product mass of atoms is finite

more than one third of MPs sit next to each other

most people procrastinate

the total area of Europe is greater than 5,000,000 km2

the highest mountain in Peru is the Huascaran

the average height of men in France is 180 cm

less than one fifth of Brazilians like cricket

the product mass of atoms is finite

more than one third of MPs sit next to each other

most people procrastinate

Motivation (ctd.) III

the total area of Europe is greater than 5,000,000 km2

the highest mountain in Peru is the Huascaran

the average height of men in France is 180 cm

less than one fifth of Brazilians like cricket

the product mass of atoms is finite

more than one third of MPs sit next to each other

most people procrastinate

Motivation (ctd.) IV

the total area of Europe is greater than 5,000,000 km2

the highest mountain in Peru is the Huascaran

the average height of men in France is 180 cm

less than one fifth of Brazilians like cricket

the product mass of atoms is finite

more than one third of MPs sit next to each other

most people procrastinate

Outline

English Generalized Quantifiers [BC80]

Definition (Generalized Quantifier)

Given $\mathcal{I}$, a generalized quantifier Q of type $(k_1, \ldots, k_n)$ is a relation of tuples $(R_1, \ldots, R_n)$ s.t., for $1 \leq i \leq k$, $R_i \subseteq \Delta^{k_i}$.

- English generalized quantifiers are realized by **Det**s and **NP**s
- They state relations that hold over properties in a model

$$
\begin{aligned}
[\![\text{no}]\!] &= \{(A, B) \subseteq \Delta \times \Delta \mid A \cap B = \emptyset\} \\
[\![\text{every}]\!] &= \{(A, B) \subseteq \Delta \times \Delta \mid A \subseteq B\} \\
[\![\text{at least } k]\!] &= \{(A, B) \subseteq \Delta \times \Delta \mid \#(A \cap B) \geq k\} \\
[\![\text{some}]\!] &= \{(A, B) \subseteq \Delta \times \Delta \mid A \cap B \neq \emptyset\}
\end{aligned}
$$

FOL quantifiers of type (1,1)

Proportional and Aggregate Quantifiers

$$
\begin{array}{rcl}
[\![\text{the number of}]\!] & = & \{(A,B) \subseteq \Delta \times \Delta \mid \mathbf{count}(A) \in B\} \\
[\![\text{the average } \beta \text{ of}]\!] & = & \{(A,B) \subseteq \Delta \times \Delta \mid \mathbf{avg}(\beta(A)) \in B\} \\
[\![\text{the total } \beta \text{ of}]\!] & = & \{(A,B) \subseteq \Delta \times \Delta \mid \mathbf{sum}(\beta(A)) \in B\} \\
[\![\text{the } \beta\text{-est}]\!] & = & \{(A,B) \subseteq \Delta \times \Delta \mid \mathbf{argmax}(\beta(A)) \in B\} \\
[\![\text{the product } \beta \text{ of}]\!] & = & \{(A,B) \subseteq \Delta \times \Delta \mid \mathbf{prod}(\beta(A)) \in B\}
\end{array}
$$

Aggregate quantifiers [Tho10] of type (1,1)

$$
\begin{array}{rcl}
[\![\text{most}]\!] & = & \{(A,B) \subseteq \Delta \times \Delta \mid \#(A \cap B) \geq \#(A \setminus B)\} \\
[\![\text{more than } n/k \text{ of}]\!] & = & \{(A,B) \subseteq \Delta \times \Delta \mid \#(A \cap B) \geq n/k \cdot \#(A)\}
\end{array}
$$

Proportional quantifiers of type (1,1)

Definition (L-Expressibility)

A quantifier Q of type $(k_1, \ldots, k_n)$ is *expressible* in logic L iff there exists a formula $\overline{Q}(R_1, \ldots, R_n)$, with R_i a relation symbol of arity k_i, for $1 \leq i \leq k$, such that, for all models $\mathcal{I}$,

$$Q = \{(R_1^{\mathcal{I}}, \ldots, R_n^{\mathcal{I}}) \subseteq \Delta^{k_1} \times \cdots \times \Delta^{k_n} \mid \mathcal{I} \models \overline{Q}(R_1, \ldots, R_n)\}$$

$$[\![\text{no}]\!] = \{(A^{\mathcal{I}}, B^{\mathcal{I}}) \subseteq \Delta \times \Delta \mid \mathcal{I} \models \forall x(A(x) \Rightarrow \neg B(x))\}$$

$$[\![\text{some}]\!] = \{(A^{\mathcal{I}}, B^{\mathcal{I}}) \subseteq \Delta \times \Delta \mid \mathcal{I} \models \exists x(A(x) \wedge B(x))\}$$

Q: Are proportional and aggregate quantifiers more expressive or complex than FOL quantifiers?

Expressiveness: $\mathbf{argmin}(\cdot), \mathbf{argmax}(\cdot)$

Theorem

If we consider Δ ordered by $\leq$ then the functions $\mathbf{argmin}(\cdot)$ and $\mathbf{argmax}(\cdot)$ are FOL-expressible

- Indeed, for all $\mathcal{I}$,

$$\mathcal{I} \models c \approx \mathbf{argmax}(P)$$

iff

$$\mathcal{I} \models \exists! x \forall y (P(x) \land P(y) \land x \geq y \land x \approx c)$$

Theorem

If we order the domain, the quantifier "the β-est" (and comparatives) is FOL-expressible

Expressiveness: $\mathbf{count}(\cdot), \mathbf{sum}(\cdot), \mathbf{prod}(\cdot)$

Theorem

If we consider $\mathbf{Rat} = (\mathbb{Q}; +, \times; \geq)$ *(ordered field of the reals) to hold, then:*

1. $\mathbf{prod}(\cdot)$ *and* $\mathbf{avg}(\cdot)$ *are definable in terms of* $\mathbf{sum}(\cdot)$ *and* $\mathbf{count}(\cdot)$
2. $\mathbf{sum}(\cdot)$ *is definable in terms of* $\mathbf{count}(\cdot)$
3. *the quantifier "most" is definable in terms of "the number of"*

- Recall: $[\![\text{most}]\!] = \{(A, B \subseteq \Delta \times \Delta \mid \mathbf{count}(A \cap B) \geq \mathbf{count}(A \setminus B)\}$

Theorem

Aggregate quantifiers are not FOL-expressible

- The generalized quantifier "most" is not FOL-expressible [BC80]

Semantic Complexity [PH10]

Definition (Semantic Complexity)

Given model $\mathcal{I}$, the semantic complexity of quantifier Q expressible by $\overline{Q}(A, B)$ is defined as the cost of computing $\mathcal{I}, \gamma \models \overline{Q}(A, B)$, for some $\gamma \in \Delta^{FV(\overline{Q}(A,B))}$

- Computational cost $=$ computational complexity
- We measure cost only in $\#(\Delta)$: data complexity

- If data complexity:
  1. is at most in P: Q tractable
  2. lies beyond P: Q intractable

Remark

We consider the (simple) hierarchy: $\mathrm{AC}^0 \subseteq \mathrm{L} \subseteq \mathrm{P} \subseteq \mathrm{NP}$-complete $\subseteq \mathrm{NP}$

Tractable Quantifier Complexity I

Quantifier	D.C.
some	AC^0
every	AC^0
at least k	AC^0
more than k	AC^0
exactly k	AC^0
the α-est	AC^0
the total α of	L
the number of	L
the average α of	L
the product α of	L
most	L
more than p/k of	L

$\Rightarrow$ FOL quantifiers

Tractable Quantifier Complexity II

Quantifier	D.C.
some	AC^0
every	AC^0
at least k	AC^0
more than k	AC^0
exactly k	AC^0
the α-est	AC^0
the total α of	L
the number of	L
the average α of	L
the product α of	L
most	L
more than p/k of	L

$\Rightarrow$ Beyond FOL

Ramsey Quantifiers [Szy10]

Definition (Ramseyfication)

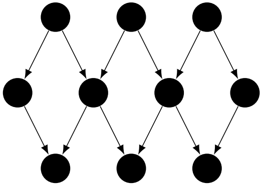
The Ramseyfication of Q of type (1,1) is the quantifier of type (1,2)

$$R_Q = \{(A, R) \subseteq \Delta \times \Delta^2 \mid \text{exists } X \subseteq A \text{ s.t.} (A, X) \in Q \text{ and for all } x, y \in X, (x, y) \in R\}$$

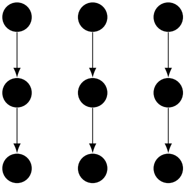
- "Says" that the As that fall under Q are R-connected
- Are conveyed in English by the reciprocal **NP** "each other"
- Can be used to express graph properties such as the existence of cliques
- They are not FOL expressible

Ramsey Quantifiers Example

more than one third of PMs sit next to each other



model $\mathcal{I}_1$

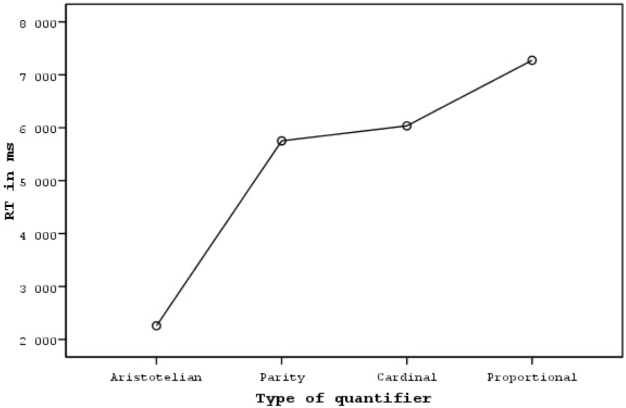


model $\mathcal{I}_2$

Ramsey Quantifiers [Szy10] (ctd.)

Quantifier	**D.C.**
some + each other	P
every + each other	P
exactly k + each other	P
most + each other	P
at least k + each other	NP-complete* (P)
at least k + each other	NP-complete* (P)
more than k + each other	NP-complete* (P)
more than p/k of + each other	NP-complete

Answer Time and Complexity [Szy09]



Power Law Distributions [Bar09]

Definition (Power law)

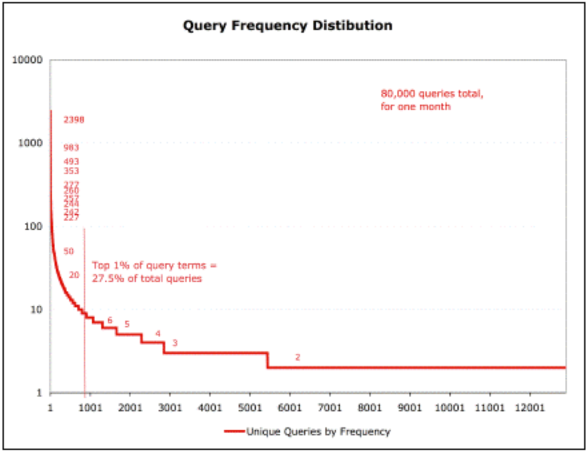
We say that a random variable X of outcomes $x_1, \ldots, x_k$ follows a power law or Zipf distribution if $\leq 20\%$ of its outcomes concentrate $\geq 80\%$ of its probability mass. This relation is described by the equation:

$$P(x) \sim \frac{b}{\mathit{rank}(x)^m}$$

- We want to know if quantifier distribution $P(Q)$ is power-law correlated to quantifier expressiveness/complexity:

$$P(Q) \sim \frac{b}{\mathit{comp}(Q)^m}$$

Power Law Example



Corpora

Corpus	Size	Domain	Type
Brown	19,741 sentences	Open (news)	Declarative
Geoquery	364 questions	Geographical	Interrogative
Clinical ques.	12,189 questions	Clinical	Interrogative
TREC 2008	436 questions	Open	Interrogative

Remark

Corpora of different types and domains and approx. 1,000,000 words (cumulatively)

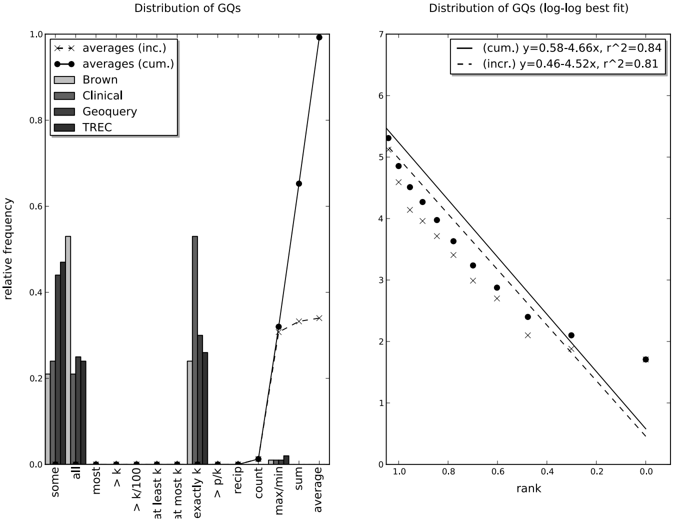
Power Laws and Log-Log Regressions

- We can transform power laws to linear models via logarithmic scaling

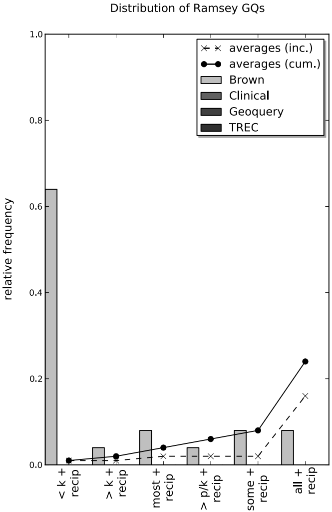
$$
\begin{aligned}
y &= b/x^m \\
&\Leftrightarrow \\
\log_{10}(y) &= \log_{10}(b) - m \cdot \log_{10}(x)
\end{aligned}
$$

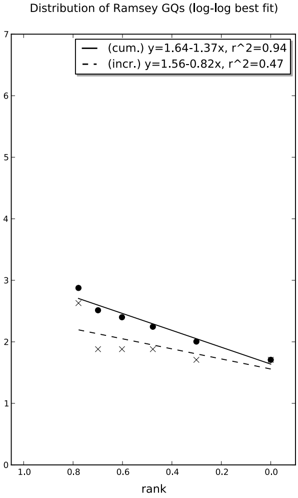
- We can estimate b and m from a sample $\mathcal{S}$ via linear regression
- If R^2 coefficient is high $\Rightarrow$ $\mathcal{S}$ power law distributed

Quantifier Distribution (all)



Ramsey Quantifier Distribution





Test Statistics

skewness	**Recip. GQs**	**GQs**
skew. value	1.76	1.98

χ^2-test	**Recip. GQs**	**GQs**
χ^2 value	530.81	183815415173.11
p value, d.f.	1.78, 5	0.0, 13

R^2-coeff.	**Recip. GQs**	**GQs**
Power law $fr(Q)$	$36.00/rk(Q)^{0.82}$	$2.88/rk(Q)^{4.52}$
R^2 coeff.	0.47	0.81

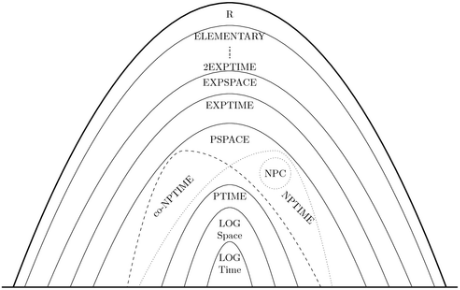
Remark

Power laws of mean relative frequency

Conclusions

1. We have studied the distribution of FOL, proportional and aggregate generalized quantifiers in corpora
2. It may seem that their distributions is skewed towards low complexity quantifiers
3. The skewed distribution is consistent with cognitive experiments [BSS11]
4. We have considered if such distribution can be modeled by a power law

Thank you :-)



http://www.inf.unibz.it/~cathorne

References I

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

John Barwise and Robin Cooper.
Generalized quantifiers and natural language.
Linguistics and Philosophy, 4(2):159–219, 1980.

Oliver Bott, Fabian Schlotterbeck, and Jakub Szymanik.
Interpreting tractable versus intractable reciprocal sentences.
In *Proceedings of the 3rd Intenational Conference in Computational Semantics (IWCS 2011)*, 2011.

Ian Pratt-Hartmann.
Computational complexity in natural language.
In *Handbook of Computational Linguistics and Natural Language Processing*, chapter 2, pages 43–73. Wiley-Blackwell, 2010.

References II

Jakub Szymanik.
Quantifiers in Time and Space.
Institute for Logic, Language and Computation, 2009.

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

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

Aggregations [Tho10]

Definition (Aggregation Function)

An aggregate function is a is a function that takes as argument a group G and returns a number $n \in \mathbb{Q}$, viz.,

$$\begin{array}{ccc} \mathbf{count}(G) & \mathbf{sum}(G) & \mathbf{argmin}(G) \\ \mathbf{avg}(G) & \mathbf{prod}(G) & \mathbf{argmax}(G) \end{array}$$

- They require models with a ordered numerical domain $N \subseteq \Delta$, with N a finite subset of $\mathbb{Q}$
- The argument group G is built via, possibly, metric attributes $\beta(\cdot)$

Tractable Quantifiers

Theorem

The semantic (data) complexity of FOL quantifiers is in AC^0

- Known result from FOL finite model theory

Theorem

The semantic (data) complexity of aggregate quantifiers (and proportional quantifiers) is in L

- One can design a sound an complete algorithm $\textsc{Ans}_{\alpha}(\mathcal{I}, \overline{Q}(A, B))$ for solving $\mathcal{I} \models \overline{Q}(A, B)$ that runs in space $O(\log \#(\Delta))$

procedure $\textsc{Ans}_\alpha(Q(\alpha(\beta(P))), \mathcal{I})$
    $\varphi(x)_P \leftarrow \textsc{Core}(Q(\alpha(\beta(P))))$; ▷ compute core
    $s \leftarrow 0$; $a \leftarrow 0$; $n \leftarrow 0$; $p \leftarrow 0$; ▷ initialize
    for $\gamma \in \mathit{Sat}_{\mathcal{I}}(\varphi(x))$ **do** ▷ $\mathit{Sat}_{\mathcal{I}}(\varphi(x)) = \{\gamma \mid \mathcal{I}, \gamma \models \varphi(x)\}$
        $n \leftarrow n + 1$; $s \leftarrow s + \beta(\gamma(x))$; ▷ update 1
        $a \leftarrow \frac{s}{n}$; $p \leftarrow p \times \beta(\gamma(x))$; ▷ update 2
        if $\alpha = \mathbf{count}$ and $Q(n)$ **then** ▷ test 1
            return true;
        else
            if $\alpha = \mathbf{avg}$ and $Q(a)$ **then** ▷ test 2
                return true;
            else
                if $\alpha = \mathbf{sum}$ and $Q(s)$ **then** ▷ test 3
                    return true;
                else
                    if $\alpha = \mathbf{prod}$ and $Q(p)$ **then** ▷ test 4
                        return true;
                    end if
                end if
            end if
        end if
    end for
    return false; ▷ false if all tests fail
end procedure

Linear Regression (Reminder)

A linear regression model has the form:

$$Y = \Theta X$$

with parameters $\Theta = (m, b)^T$ (a gradient and an intercept)

The least squares method infers from training sample $\mathcal{S} = \{(x_i, y_i)\}_{i \in [1,n]}$ the model whose parameters Θ^*:

$$\Theta^* = \arg\min_{\Theta} J(\Theta) = \arg\min_{\Theta} \sum_{i=1}^{n} (y_i - \Theta(x_i))^2$$

minimize square error

The R^2 coefficient provides a measure of confidence in $Y = \Theta^* X$:

$$R^2 = \frac{Var(\Theta^* X)}{Var(Y)}$$

Ramsey and non-Ramsey (raw)

Corpus	$>k+$ recip	$>p/k+$ recip	most+ recip	some+ recip	all+ recip	$<k+$ recip
Brown	1	1	2	2	2	16
TREC	0	0	0	0	0	0
Geo	0	0	0	0	0	0
Clin. qs.	0	0	0	0	0	0
total	1	1	2	2	2	16

Corpus	$\geq k$	$\leq k$	most	$>k$	$>p/k$	recip.	$>k\%$	sum	cnt	avg	max,min	all	k	some
Brown	192	4	1532	540	38	101	2	1	354	17	4368	202587	90811	81693
TREC	0	0	0	0	0	0	0	0	0	0	13	192	490	222
Geo	2	0	0	0	0	0	0	0	1	0	18	380	447	660
Clin. qs.	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