Quantifier Distribution and Semantic Complexity

Camilo Thorne\textsuperscript{1} \hspace{1cm} Jakub Szymanik\textsuperscript{2}

\textsuperscript{1}KRDB Research Centre for Knowledge and Data
cthorne@inf.unibz.it
http://www.inf.unibz.it/~cathorne

\textsuperscript{2}Institute for Logic, Language and Computation
jakub.szymanik@gmail.com
http://www.jakubszymanik.com/

TbiLLC2013, Sep 26, Tbilisi
Motivation

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 km\(^2\)

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 km\(^2\)

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
Outline

1. Background

2. Generalized Quantifiers
   - Reminder
   - Expressive Power

3. Semantic Complexity
   - Tractable Quantifiers
   - Intractable Quantifiers

4. Quantifier Distribution
   - Power Laws
   - Corpora
   - Results

5. Conclusions

6. References

7. Appendix
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

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

FOL quantifiers of type $(1,1)$
Proportional and Aggregate Quantifiers

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

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

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

Proportional quantifiers of type (1,1)
L-Expressibility

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 $\mathcal{I}$,

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

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

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

Q: Are proportional and aggregate quantifiers more expressive or complex than FOL quantifiers?
Expressiveness: \( \text{argmin}(\cdot), \text{argmax}(\cdot) \)

**Theorem**

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

- Indeed, for all \( \mathcal{I} \),

\[
\mathcal{I} \models c \approx \text{argmax}(P) \quad \text{iff} \quad \mathcal{I} \models \exists! x \forall y (P(x) \land P(y) \land x \geq y \land x \approx c)
\]

**Theorem**

If we order the domain, the quantifier “the \( \beta \)-est” (and comparatives) is FOL-expressible
**Theorem**

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

1. $\text{prod}(\cdot)$ and $\text{avg}(\cdot)$ are definable in terms of $\text{sum}(\cdot)$ and $\text{count}(\cdot)$
2. $\text{sum}(\cdot)$ is definable in terms of $\text{count}(\cdot)$
3. The quantifier “most” is definable in terms of “the number of”

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

**Theorem**

Aggregate quantifiers are not FOL-expressible

▷ The generalized quantifier “most” is not FOL-expressible [BC80]
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: $AC^0 \subseteq L \subseteq P \subseteq NP$-complete $\subseteq NP$
### Tractable Quantifier Complexity I

<table>
<thead>
<tr>
<th>Quantifier</th>
<th>D.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>some</td>
<td>$\text{AC}^0$</td>
</tr>
<tr>
<td>every</td>
<td>$\text{AC}^0$</td>
</tr>
<tr>
<td>at least $k$</td>
<td>$\text{AC}^0$</td>
</tr>
<tr>
<td>more than $k$</td>
<td>$\text{AC}^0$</td>
</tr>
<tr>
<td>exactly $k$</td>
<td>$\text{AC}^0$</td>
</tr>
<tr>
<td>the $\alpha$-est</td>
<td>$\text{AC}^0$</td>
</tr>
<tr>
<td>the total $\alpha$ of</td>
<td>$\text{L}$</td>
</tr>
<tr>
<td>the number of</td>
<td>$\text{L}$</td>
</tr>
<tr>
<td>the average $\alpha$ of</td>
<td>$\text{L}$</td>
</tr>
<tr>
<td>the product $\alpha$ of</td>
<td>$\text{L}$</td>
</tr>
<tr>
<td>most</td>
<td>$\text{L}$</td>
</tr>
<tr>
<td>more than $p/k$ of</td>
<td>$\text{L}$</td>
</tr>
</tbody>
</table>

$\Rightarrow$ FOL quantifiers
### Quantifier Complexity

<table>
<thead>
<tr>
<th>Quantifier</th>
<th>D.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>some</td>
<td>$AC^0$</td>
</tr>
<tr>
<td>every</td>
<td>$AC^0$</td>
</tr>
<tr>
<td>at least $k$</td>
<td>$AC^0$</td>
</tr>
<tr>
<td>more than $k$</td>
<td>$AC^0$</td>
</tr>
<tr>
<td>exactly $k$</td>
<td>$AC^0$</td>
</tr>
<tr>
<td>the $\alpha$-est</td>
<td>$AC^0$</td>
</tr>
</tbody>
</table>

$\Rightarrow$ Beyond FOL

<table>
<thead>
<tr>
<th></th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>the total $\alpha$ of</td>
<td></td>
</tr>
<tr>
<td>the number of</td>
<td></td>
</tr>
<tr>
<td>the average $\alpha$ of</td>
<td></td>
</tr>
<tr>
<td>the product $\alpha$ of</td>
<td></td>
</tr>
<tr>
<td>most</td>
<td>L</td>
</tr>
<tr>
<td>more than $p/k$ of</td>
<td>L</td>
</tr>
</tbody>
</table>
Ramsey Quantifiers [Szy10]

**Definition (Ramseyfication)**

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

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

- "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
more than one third of PMs sit next to each other

model $\mathcal{I}_1$  

model $\mathcal{I}_2$
### Quantifier

<table>
<thead>
<tr>
<th>Quantifier</th>
<th>D.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>some + each other</td>
<td>P</td>
</tr>
<tr>
<td>every + each other</td>
<td>P</td>
</tr>
<tr>
<td>exactly $k$ + each other</td>
<td>P</td>
</tr>
<tr>
<td>most + each other</td>
<td>P</td>
</tr>
<tr>
<td>at least $k$ + each other</td>
<td>NP-complete* (P)</td>
</tr>
<tr>
<td>at least $k$ + each other</td>
<td>NP-complete* (P)</td>
</tr>
<tr>
<td>more than $k$ + each other</td>
<td>NP-complete* (P)</td>
</tr>
<tr>
<td>more than $p/k$ of + each other</td>
<td>NP-complete</td>
</tr>
</tbody>
</table>
Answer Time and Complexity [Szy09]
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:

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

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

$$P(Q) \sim \frac{b}{\text{comp}(Q)^m}$$
Power Law Example

Query Frequency Distribution

80,000 queries total, for one month

Top 1% of query terms = 27.5% of total queries

Unique Queries by Frequency

©2006 Search Tools Consulting
### Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
<th>Domain</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>19,741 sentences</td>
<td>Open (news)</td>
<td>Declarative</td>
</tr>
<tr>
<td>Geoquery</td>
<td>364 questions</td>
<td>Geographical</td>
<td>Interrogative</td>
</tr>
<tr>
<td>Clinical ques.</td>
<td>12,189 questions</td>
<td>Clinical</td>
<td>Interrogative</td>
</tr>
<tr>
<td>TREC 2008</td>
<td>436 questions</td>
<td>Open</td>
<td>Interrogative</td>
</tr>
</tbody>
</table>

**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

\[ y = \frac{b}{x^m} \]

\[ \iff \]

\[ \log_{10}(y) = \log_{10}(b) - m \cdot \log_{10}(x) \]

- We can estimate \( b \) and \( m \) from a sample \( S \) via linear regression

- If \( R^2 \) coefficient is high \( \Rightarrow S \) power law distributed
Quantifier Distribution (all)

Distribution of GQs

- averages (inc.)
- averages (cum.)
- Brown
- Clinical
- Geoquery
- TREC

Distribution of GQs (log-log best fit)

- (cum.) y = 0.58 - 4.66x, r^2 = 0.84
- (incr.) y = 0.46 - 4.52x, r^2 = 0.81
Ramsey Quantifier Distribution

The graphs show the distribution of Ramsey GQs (Generalized Quantifiers) and their log-log best fit. The distribution plots illustrate the frequency of different quantifiers, with axes labeled as 'relative frequency' and 'rank'. The graphs are labeled as follows:

- **Distribution of Ramsey GQs**: Graph showing the frequency of different quantifiers.
- **Distribution of Ramsey GQs (log-log best fit)**: Graph showing the log-log best fit for the frequency of different quantifiers.

Key points:
- The graphs use markers and lines to represent different quantifiers and their frequencies.
- The axes are labeled as 'relative frequency' and 'rank'.
- The plots include annotations for the best fit lines with their respective regression coefficients.

The document references C. Thorne, J. Szymanik (KRDB, ILLC) and includes information on Quantifier Distribution at TbiLLC2013, Sep 26, Tbilisi.
## Test Statistics

<table>
<thead>
<tr>
<th>skewness</th>
<th>Recip. GQs</th>
<th>GQs</th>
</tr>
</thead>
<tbody>
<tr>
<td>skew. value</td>
<td>1.76</td>
<td>1.98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\chi^2$-test</th>
<th>Recip. GQs</th>
<th>GQs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$ value</td>
<td>530.81</td>
<td>183815415173.11</td>
</tr>
<tr>
<td>$p$ value, d.f.</td>
<td>1.78, 5</td>
<td>0.0, 13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$R^2$-coeff.</th>
<th>Recip. GQs</th>
<th>GQs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power law $fr(Q)$</td>
<td>$36.00/rk(Q)^{0.82}$</td>
<td>$2.88/rk(Q)^{4.52}$</td>
</tr>
<tr>
<td>$R^2$ coeff.</td>
<td>0.47</td>
<td>0.81</td>
</tr>
</tbody>
</table>

## 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 :-)
References I


Jakub Szymanik.  
*Quantifiers in Time and Space.*  

Jakub Szymanik.  
Computational complexity of polyadic lifts of generalized quantifiers in natural language.  

Camilo Thorne.  
*Query Answering over Ontologies Using Controlled Natural Languages.*  
Aggregations [Tho10]

Definition (Aggregation Function)

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

\[
\begin{align*}
\text{count}(G) & \quad \text{sum}(G) & \quad \text{argmin}(G) \\
\text{avg}(G) & \quad \text{prod}(G) & \quad \text{argmax}(G)
\end{align*}
\]

- 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 $\mathbf{AC}^0$

- Known result from FOL finite model theory

Theorem

The semantic (data) complexity of aggregate quantifiers (and proportional quantifiers) is in $\mathbf{L}$

- One can design a sound and complete algorithm $\text{ANS}_\alpha(\mathcal{I}, \overline{Q}(A, B))$ for solving $\mathcal{I} \models \overline{Q}(A, B)$ that runs in space $O(\log \#(\Delta))$
Answering Aggregations \( (O(\log \#(\Delta)) \text{ Space}) \)

1: **procedure** \( \text{ANS}_\alpha(Q(\alpha(\beta(P))), \mathcal{I}) \)  
2: \( \varphi(x)_P \leftarrow \text{CORE}(Q(\alpha(\beta(P)))) \);  
3: \( s \leftarrow 0; a \leftarrow 0; n \leftarrow 0; p \leftarrow 0; \)  
4: **for** \( \gamma \in \text{Sat}_\mathcal{I}(\varphi(x)) \) **do**  
5: \( n \leftarrow n + 1; s \leftarrow s + \beta(\gamma(x)) \);  
6: \( a \leftarrow \frac{s}{n}; p \leftarrow p \times \beta(\gamma(x)) \);  
7: **if** \( \alpha = \text{count} \) and \( Q(n) \) **then** \( \text{return true;} \)  
8: **else**  
9: **if** \( \alpha = \text{avg} \) and \( Q(a) \) **then** \( \text{return true;} \)  
10: **else**  
11: **if** \( \alpha = \text{sum} \) and \( Q(s) \) **then** \( \text{return true;} \)  
12: **else**  
13: **if** \( \alpha = \text{prod} \) and \( Q(p) \) **then** \( \text{return true;} \)  
14: **end if**  
15: **end if**  
16: **end if**  
17: **end if**  
18: **end for**  
19: **return** false;  
20: **end procedure**  

\( \triangleright \) compute core  
\( \triangleright \) initialize  
\( \triangleright \) Sat\(_\mathcal{I}(\varphi(x)) = \{\gamma | \mathcal{I}, \gamma \models \varphi(x)\} \)  
\( \triangleright \) update 1  
\( \triangleright \) update 2  
\( \triangleright \) test 1  
\( \triangleright \) test 2  
\( \triangleright \) test 3  
\( \triangleright \) test 4  
\( \triangleright \) false if all tests fail
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 $S = \{(x_i, y_i)\}_{i \in [1,n]}$ the model whose parameters $\Theta^*$:

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

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)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>(&gt; k^+) reciprocity</th>
<th>(&gt; \frac{p}{k^+}) reciprocity</th>
<th>most (&gt; k) reciprocity</th>
<th>some (&gt; k) reciprocity</th>
<th>all (&gt; k) reciprocity</th>
<th>(&lt; k^+) reciprocity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>TREC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Geo</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Clin. qs.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>total</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Corpus</th>
<th>(\geq k)</th>
<th>(\leq k)</th>
<th>most (&gt; k)</th>
<th>(&gt; \frac{p}{k}) reciprocity</th>
<th>(&gt; k)% sum</th>
<th>cnt</th>
<th>avg</th>
<th>max, min</th>
<th>all</th>
<th>(k)</th>
<th>some</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>192</td>
<td>4</td>
<td>1532</td>
<td>540</td>
<td>38</td>
<td>101</td>
<td>2</td>
<td>1</td>
<td>354</td>
<td>17</td>
<td>4368</td>
</tr>
<tr>
<td>TREC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Geo</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>Clin. qs.</td>
<td>12</td>
<td>0</td>
<td>28</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>2</td>
<td>889</td>
</tr>
<tr>
<td>total</td>
<td>206</td>
<td>4</td>
<td>1560</td>
<td>552</td>
<td>38</td>
<td>101</td>
<td>2</td>
<td>1</td>
<td>364</td>
<td>19</td>
<td>5288</td>
</tr>
</tbody>
</table>

C. Thorne, J. Szymanik (KRDB, ILLC)