Generalized Quantifiers in Corpora

Camilo Thorne[1] (Joint work with Jakub Szymanik[2])

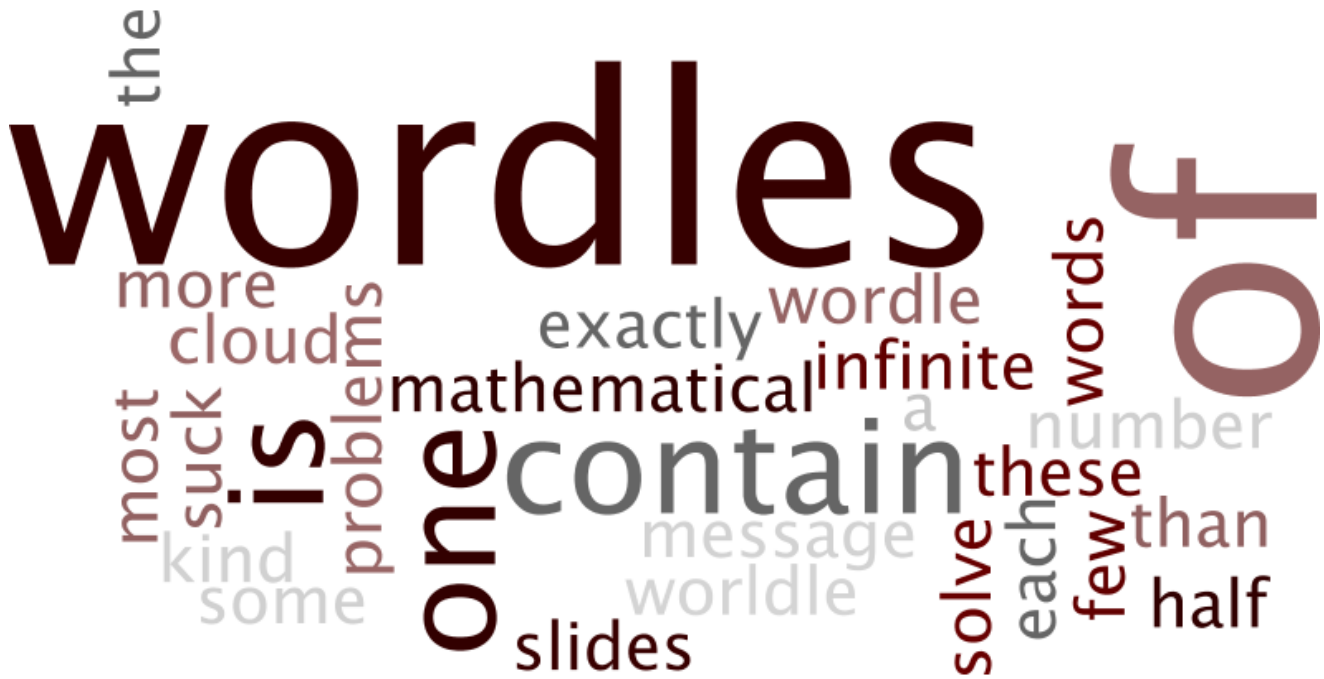
[1]DWS Group, Universität Mannheim

[2]ILLC, University of Amsterdam

DWS Talk, 27.05.2016

Outline

Motivation - Quantification in English



Motivation - Quantification in English (ctd.)

$\vdots$

each wordle is a cloud

most wordles contain some kind of message

fewer than $1/2$ of wordles suck

the number of wordles is infinite

few wordles solve mathematical problems

exactly one wordle is contained in these slides

$\vdots$

Motivation - Quantification in English (ctd.)

⋮

each wordle is a cloud

most wordles contain some kind of message

fewer than $1/2$ of wordles suck

the number of wordles is infinite

few wordles solve mathematical problems

exactly one wordle is contained in these slides

⋮

Motivation - Quantification in English (ctd.)

⋮

each wordle is a cloud

most wordles contain some kind of message

fewer than $1/2$ of wordles suck

the number of wordles is infinite

few wordles solve mathematical problems

exactly one wordle is contained in these slides

⋮

Motivation - Quantification in English (ctd.)

⋮

each wordle is a cloud

most wordles contain some kind of message

fewer than $1/2$ of wordles suck

the number of wordles is infinite

few wordles solve mathematical problems

exactly one wordle is contained in these slides

⋮

Formal Semantics

- Seminal work by Montague in [Mon70, Mon73]
- Work on quantifiers extended by Barwise and Cooper in [BC80]

Formal Semantics

- Seminal work by Montague in [Mon70, Mon73]
- Work on quantifiers extended by Barwise and Cooper in [BC80]

H1: The meaning of a natural language expression is given by its truth conditions

Formal Semantics

- Seminal work by Montague in [Mon70, Mon73]
- Work on quantifiers extended by Barwise and Cooper in [BC80]

H1: The meaning of a natural language expression is given by its truth conditions

H2: Truth conditions can be described with formal logic

Formal Semantics

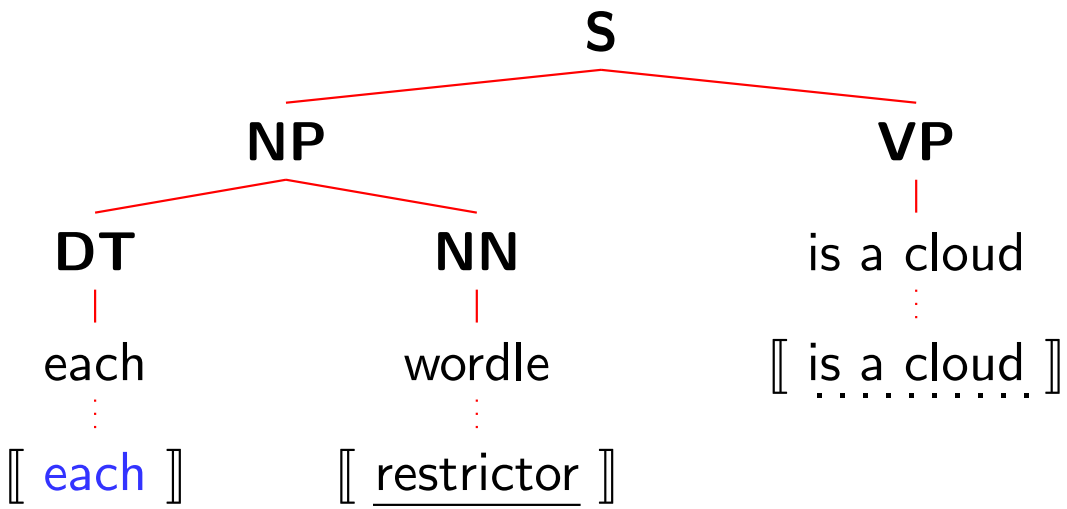
- Seminal work by Montague in [Mon70, Mon73]
- Work on quantifiers extended by Barwise and Cooper in [BC80]

H1: The meaning of a natural language expression is given by its truth conditions

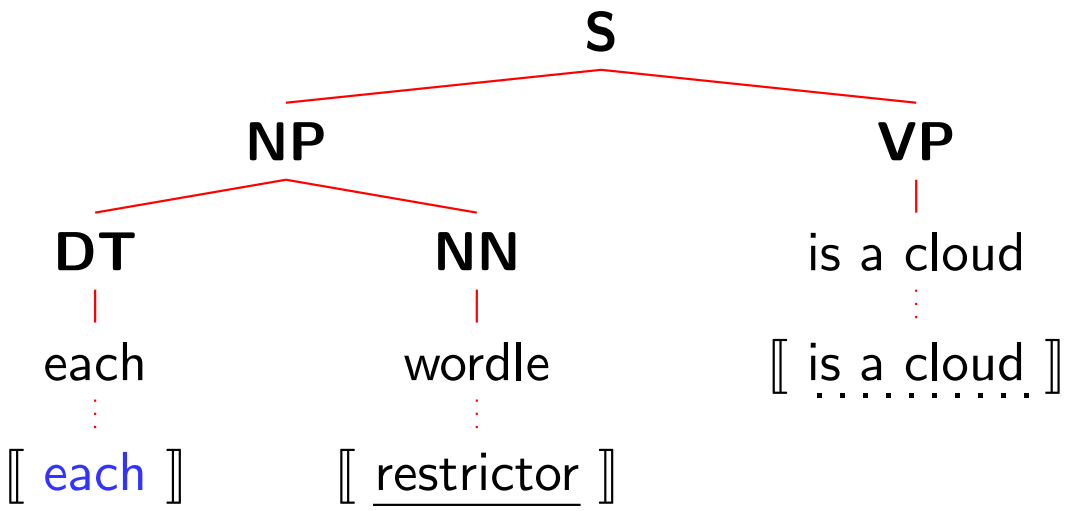
H2: Truth conditions can be described with formal logic

H3: We can model quantification via generalized quantifiers

Generalized Quantifiers

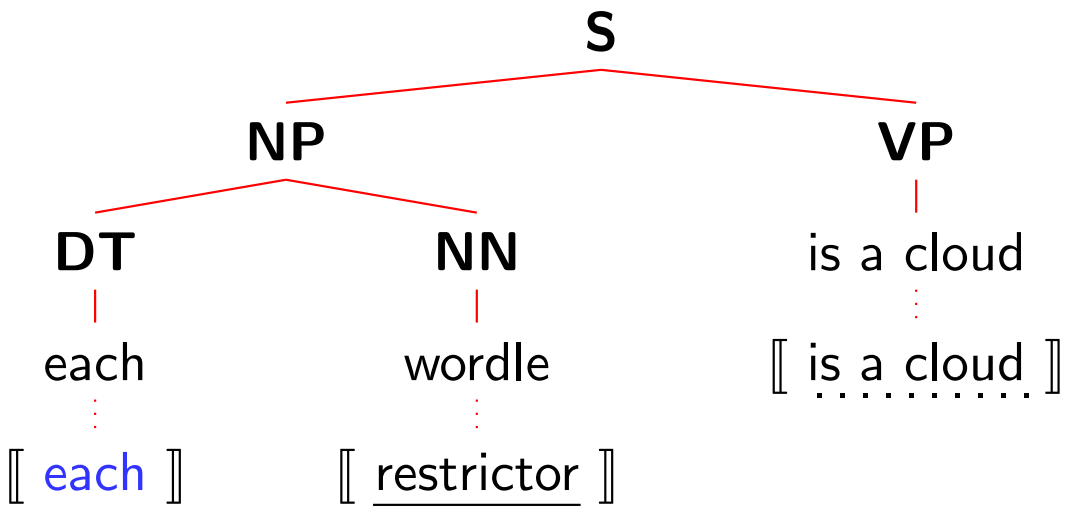


Generalized Quantifiers



$$(\llbracket \text{wordle} \rrbracket, \llbracket \text{is a cloud} \rrbracket) \in \llbracket \text{each} \rrbracket \quad \Leftrightarrow \quad \llbracket \text{wordle} \rrbracket \subseteq \llbracket \text{is a cloud} \rrbracket$$

Generalized Quantifiers



$$(\llbracket \text{wordle} \rrbracket, \llbracket \text{is a cloud} \rrbracket) \in \llbracket \text{each} \rrbracket \quad \Leftrightarrow \quad \llbracket \text{wordle} \rrbracket \subseteq \llbracket \text{is a cloud} \rrbracket$$

Definition (Generalized Quantifier)

Given the domain of discourse $\Delta = \{d_i \mid i \in \mathbb{N}\}$, a generalized quantifier Q of type $(k_1, \ldots, k_n)$ is an n-ary relation $Q \subseteq \mathcal{P}(\Delta^{k_1}) \times \cdots \times \mathcal{P}(\Delta^{k_n})$.

L-Expressibility and Complexity [Szy09]

Definition (L-Expressibility)

A quantifier Q of type $(k_1, \ldots, k_n)$ is expressible in logic L iff there exists a formula $\overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$ in L with $\overline{R}_i$ a relation symbol of arity k_i, $1 \leq i \leq n$, such that, $(R_1, \ldots, R_n) \in Q \quad \Leftrightarrow \quad (\Delta, \cdot^{\mathcal{I}}) \models \overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$

L-Expressibility and Complexity [Szy09]

Definition (L-Expressibility)

A quantifier Q of type $(k_1, \ldots, k_n)$ is expressible in logic L iff there exists a formula $\overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$ in L with $\overline{R}_i$ a relation symbol of arity k_i, $1 \leq i \leq n$, such that, $(R_1, \ldots, R_n) \in Q \quad \Leftrightarrow \quad (\Delta, \cdot^{\mathcal{I}}) \models \overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$

$$(A, B) \in [\![\text{each}]\!] \quad \Leftrightarrow \quad (\Delta, \cdot^{\mathcal{I}}) \models \forall x(\overline{A}(x) \Rightarrow \overline{B}(x)) \qquad \text{(FO-expressible)}$$

L-Expressibility and Complexity [Szy09]

Definition (L-Expressibility)

A quantifier Q of type $(k_1, \ldots, k_n)$ is expressible in logic L iff there exists a formula $\overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$ in L with $\overline{R}_i$ a relation symbol of arity k_i, $1 \leq i \leq n$, such that, $(R_1, \ldots, R_n) \in Q \quad \Leftrightarrow \quad (\Delta, \cdot^{\mathcal{I}}) \models \overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$

$$(A, B) \in [\![\text{each}]\!] \quad \Leftrightarrow \quad (\Delta, \cdot^{\mathcal{I}}) \models \forall x(\overline{A}(x) \Rightarrow \overline{B}(x)) \qquad \text{(FO-expressible)}$$

- The semantic complexity of a quantifier Q over $R_1, \ldots, R_n$ is the cost of computing $(\Delta_f, \cdot^{\mathcal{I}}) \models \overline{Q}(\overline{R}_1, \ldots, \overline{R}_n)$, Δ_f a finite subset of Δ
- We measure cost only in $\#(\Delta_f)$: data complexity

Complexity Ranking [TS15, TS13]

Q	**Semantics** $\subseteq \mathcal{P}(\Delta) \times \mathcal{P}(\Delta)$	**D.C.**		**Example**
some	$\{(A,B) \mid A \cap B \neq \emptyset\}$	L	ari	some men are happy
all	$\{(A,B) \mid A \subseteq B\}$	L		all humans are mammals
$> k$	$\{(A,B) \mid \#(A \cap B) > k\}$	L	cnt	more than 5 men are happy
$< k$	$\{(A,B) \mid \#(A \cap B) < k\}$	L		fewer than 100 violins are Stradivari
k	$\{(A,B) \mid \#(A \cap B) = k\}$	L		50 MPs voted against the war in Irak
most	$\{(A,B) \mid \#(A \cap B) > \#(A \setminus B)\}$	P	pro	most trains are safe
few	$\{(A,B) \mid \#(A \cap B) < \#(A \setminus B)\}$	P		few people are trustworthy
$> p/k$	$\{(A,B) \mid \#(A \cap B) > p \cdot (\#(A)/k)\}$	P		more than $2/3$ of planets are lifeless
$< p/k$	$\{(A,B) \mid \#(A \cap B) < p \cdot (\#(A)/k)\}$	P		less than $1/3$ of Americans are rich
p/k	$\{(A,B) \mid \#(A \cap B) = p \cdot (\#(A)/k)\}$	P		$1/3$ of Peru's population lives in Lima
$> k\%$	$\{(A,B) \mid \#(A \cap B) > k \cdot (\#(A)/100)\}$	P		more than 10% of Peruvians are poor
$< k\%$	$\{(A,B) \mid \#(A \cap B) < k \cdot (\#(A)/100)\}$	P		less than 5% of the Earth is water
$k\%$	$\{(A,B) \mid \#(A \cap B) = k \cdot (\#(A)/100)\}$	P		15% of Muslims are Shia

Complexity Ranking [TS15, TS13]

Q	**Semantics** $\subseteq \mathcal{P}(\Delta) \times \mathcal{P}(\Delta)$	**D.C.**		**Example**
some	$\{(A,B) \mid A \cap B \neq \emptyset\}$	L	ari	some men are happy
all	$\{(A,B) \mid A \subseteq B\}$	L		all humans are mammals
$> k$	$\{(A,B) \mid \#(A \cap B) > k\}$	L	cnt	more than 5 men are happy
$< k$	$\{(A,B) \mid \#(A \cap B) < k\}$	L		fewer than 100 violins are Stradivari
k	$\{(A,B) \mid \#(A \cap B) = k\}$	L		50 MPs voted against the war in Irak
most	$\{(A,B) \mid \#(A \cap B) > \#(A \setminus B)\}$	P	pro	most trains are safe
few	$\{(A,B) \mid \#(A \cap B) < \#(A \setminus B)\}$	P		few people are trustworthy
$> p/k$	$\{(A,B) \mid \#(A \cap B) > p \cdot (\#(A)/k)\}$	P		more than $2/3$ of planets are lifeless
$< p/k$	$\{(A,B) \mid \#(A \cap B) < p \cdot (\#(A)/k)\}$	P		less than $1/3$ of Americans are rich
p/k	$\{(A,B) \mid \#(A \cap B) = p \cdot (\#(A)/k)\}$	P		$1/3$ of Peru's population lives in Lima
$> k\%$	$\{(A,B) \mid \#(A \cap B) > k \cdot (\#(A)/100)\}$	P		more than 10% of Peruvians are poor
$< k\%$	$\{(A,B) \mid \#(A \cap B) < k \cdot (\#(A)/100)\}$	P		less than 5% of the Earth is water
$k\%$	$\{(A,B) \mid \#(A \cap B) = k \cdot (\#(A)/100)\}$	P		15% of Muslims are Shia

ari: Aristotelian quantifiers *cnt*: counting quantifiers *pro*: proportional quantifiers

Complexity Ranking [TS15, TS13]

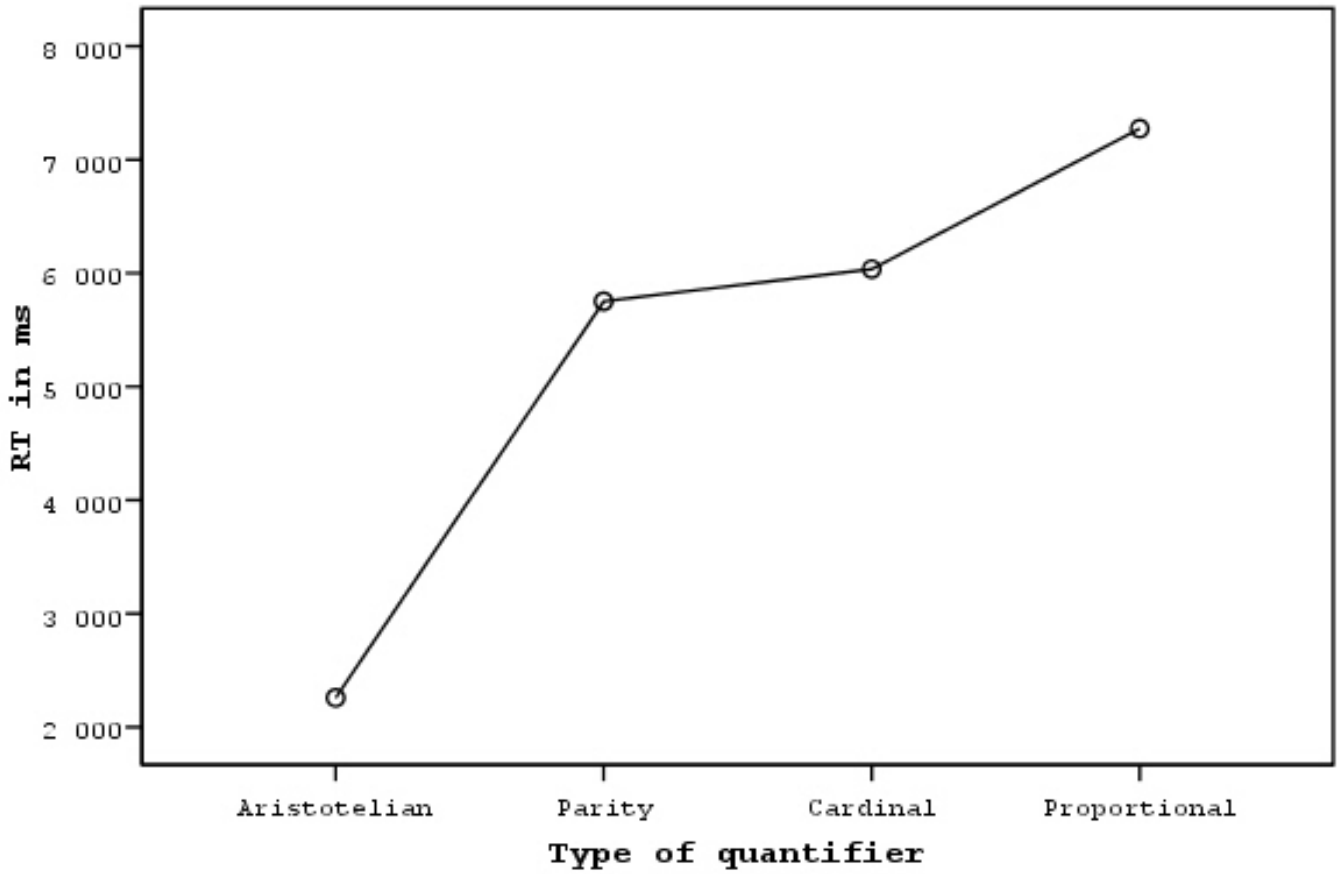
Q	**Semantics** $\subseteq \mathcal{P}(\Delta) \times \mathcal{P}(\Delta)$	**D.C.**		**Example**
some	$\{(A,B) \mid A \cap B \neq \emptyset\}$	L	ari	some men are happy
all	$\{(A,B) \mid A \subseteq B\}$	L		all humans are mammals
$> k$	$\{(A,B) \mid \#(A \cap B) > k\}$	L	cnt	more than 5 men are happy
$< k$	$\{(A,B) \mid \#(A \cap B) < k\}$	L		fewer than 100 violins are Stradivari
k	$\{(A,B) \mid \#(A \cap B) = k\}$	L		50 MPs voted against the war in Irak
most	$\{(A,B) \mid \#(A \cap B) > \#(A \setminus B)\}$	P	pro	most trains are safe
few	$\{(A,B) \mid \#(A \cap B) < \#(A \setminus B)\}$	P		few people are trustworthy
$> p/k$	$\{(A,B) \mid \#(A \cap B) > p \cdot (\#(A)/k)\}$	P		more than $2/3$ of planets are lifeless
$< p/k$	$\{(A,B) \mid \#(A \cap B) < p \cdot (\#(A)/k)\}$	P		less than $1/3$ of Americans are rich
p/k	$\{(A,B) \mid \#(A \cap B) = p \cdot (\#(A)/k)\}$	P		$1/3$ of Peru's population lives in Lima
$> k\%$	$\{(A,B) \mid \#(A \cap B) > k \cdot (\#(A)/100)\}$	P		more than 10% of Peruvians are poor
$< k\%$	$\{(A,B) \mid \#(A \cap B) < k \cdot (\#(A)/100)\}$	P		less than 5% of the Earth is water
$k\%$	$\{(A,B) \mid \#(A \cap B) = k \cdot (\#(A)/100)\}$	P		15% of Muslims are Shia

ari: Aristotelian quantifiers *cnt*: counting quantifiers *pro*: proportional quantifiers

Question

Does complexity influence quantifier distribution in (large) corpora?

Answer Time and Complexity [Szy09]



Parity: "exactly 2" *Cardinal*: all the other counting quantifiers

The WaCkY Corpus [BBFZ09]

```
<s>
Flender Flender NP      1       3       VMOD
Werke   Werke   NP      2       3       SBJ
was     be      VBD     3       0       ROOT
a       a       DT      4       7       NMOD
German  German  JJ      5       7       NMOD
shipbuilding    shipbuilding    NN      6       7       NMOD
company company NN      7       3       PRD
,       ,       ,       8       7       P
located locate  VVN     9       7       NMOD
in      in      IN      10      9       ADV
Lubeck  Lubeck  NP      11      10      PMOD
.       .       SENT    12      0       ROOT
</s>
```

	Sentences	**Tokens**	**Source**
WaCkY (Eng)	$\sim$ 43 million	$\sim$ 800 million	Wikipedia (EN, 2008)

Corpus Analysis

- We built a list of simple patterns to identify and count the quantifiers listed in **Slide 8**, i.e.,

  1. Aristotelian quantifiers: *all*, *all*
  2. counting quantifiers: k, *less (more) than* k
  3. proportional quantifiers: *most*, *few*, p/k, $k\%$, *less (more) than* p/k, *less (more) than* $k\%$

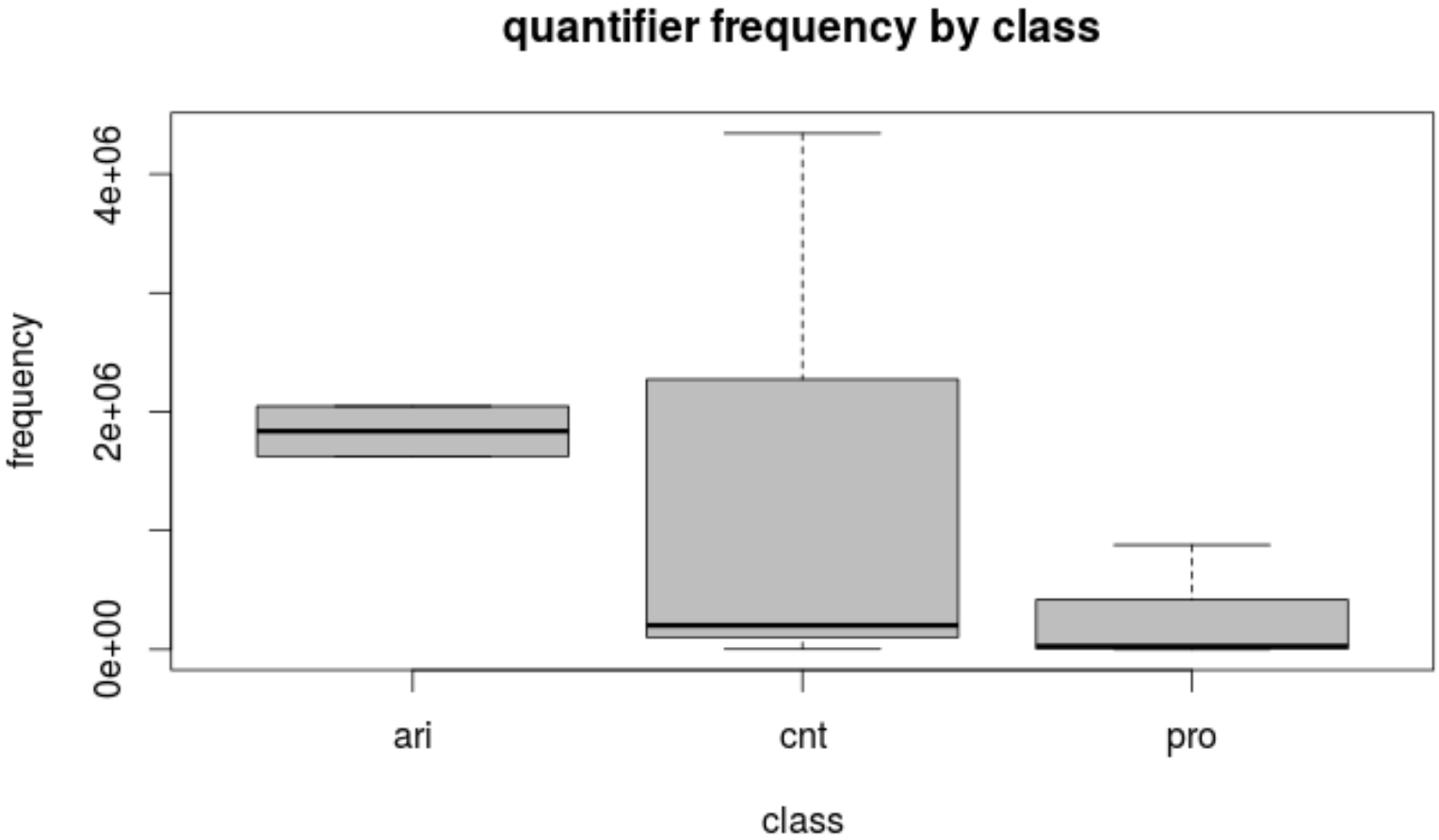
- Examples: $\begin{cases} \textit{most} = \texttt{most/dt},\ \texttt{most/jjs}\ \texttt{[a-z]\{1,12\}/nns} \\ \textit{some} = \texttt{some/det} \end{cases}$

- We seeked to understand how much their frequency is influenced by

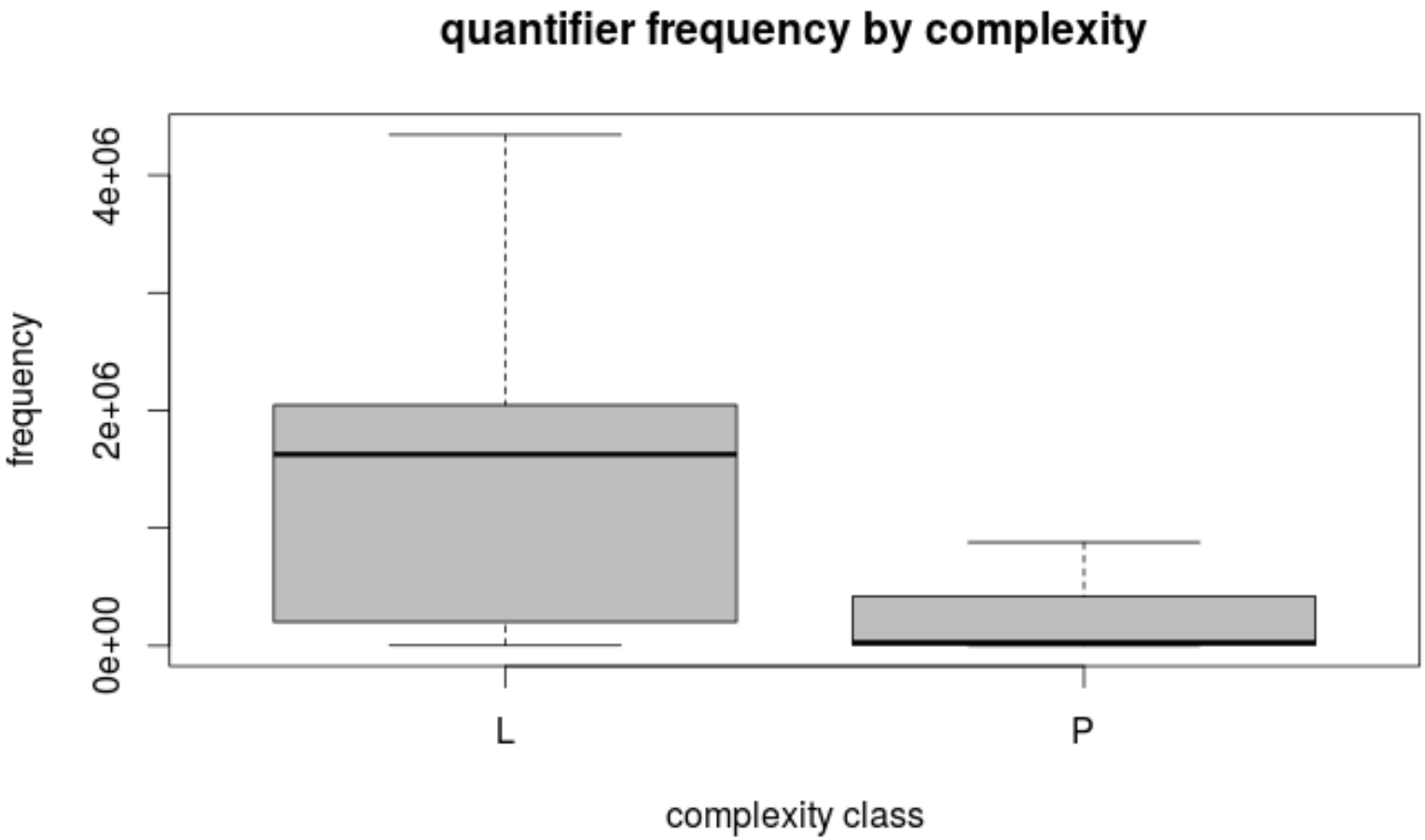
  1. length (characters, word units) $\Rightarrow$ "syntactic complexity"
  2. data complexity and/or quantifier class $\Rightarrow$ "semantic complexity"

Q	**Frequency**	**Length**	**Class**	**Characters**	**Rank**	**S.C.**
$>k$	202277	$\sim$3	*cnt*	$\sim$11	7	L
$<k$	2625	$\sim$3	*cnt*	$\sim$11	12	L
k	4342247	$\sim$1	*cnt*	$\sim$3	1	L
$>p/k$	15151	$\sim$4	*pro*	$\sim$16	9	P
$<p/k$	98	$\sim$4	*pro*	$\sim$16	13	P
p/k	589275	$\sim$2	*pro*	$\sim$8	5	P
$>k/100$	8091	$\sim$4	*pro*	$\sim$18	10	P
$<k/100$	5935	$\sim$4	*pro*	$\sim$18	11	P
$k/100$	39288	$\sim$2	*pro*	$\sim$10	8	P
few	248521	$\sim$1	*pro*	$\sim$3	6	P
most	879069	$\sim$1	*pro*	$\sim$4	4	P
all	2046388	$\sim$1	*ari*	$\sim$3	2	L
some	1626139	$\sim$1	*ari*	$\sim$4	3	L

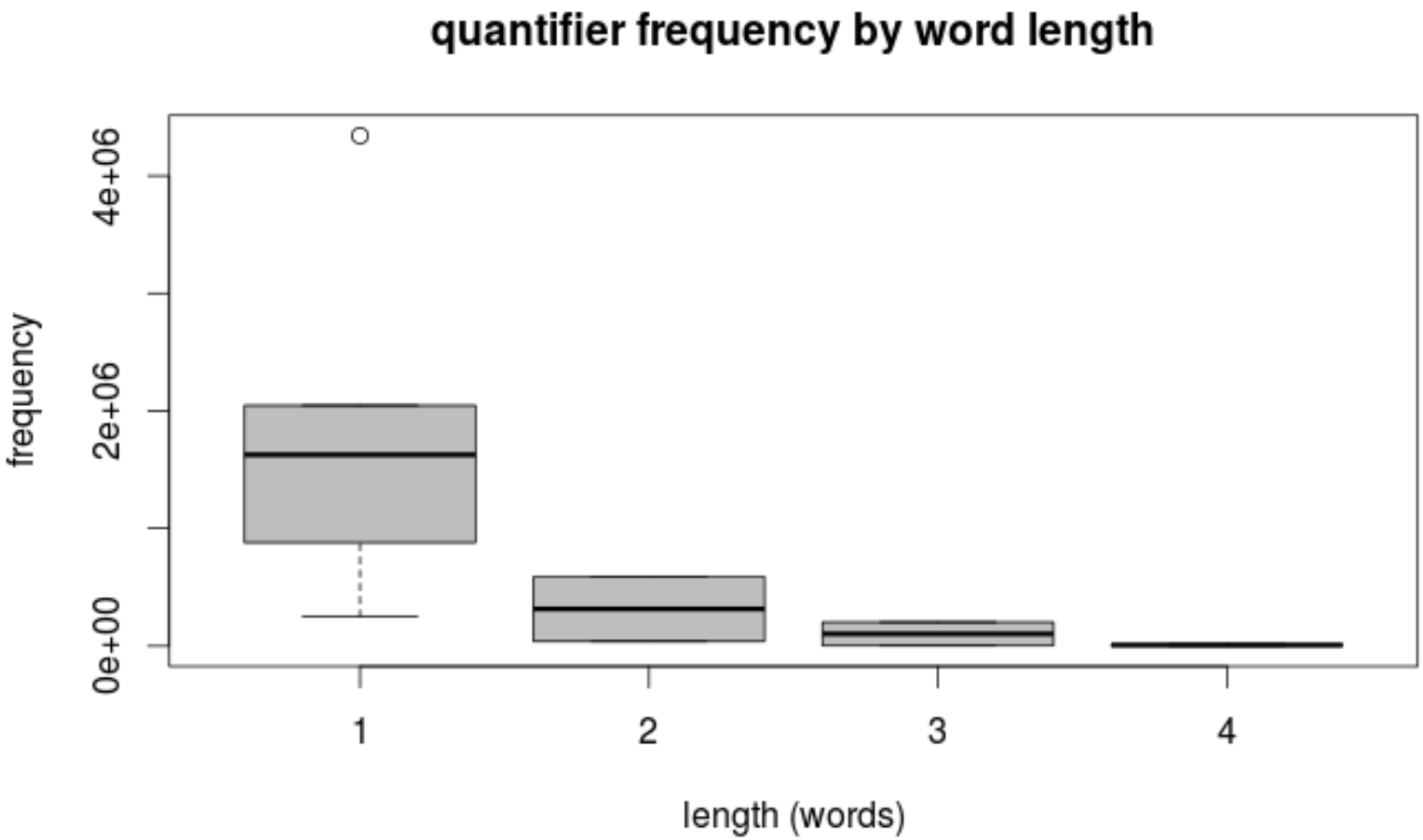
- Distribution skewed towards quantifiers of low complexity: skewness of -0.5752868



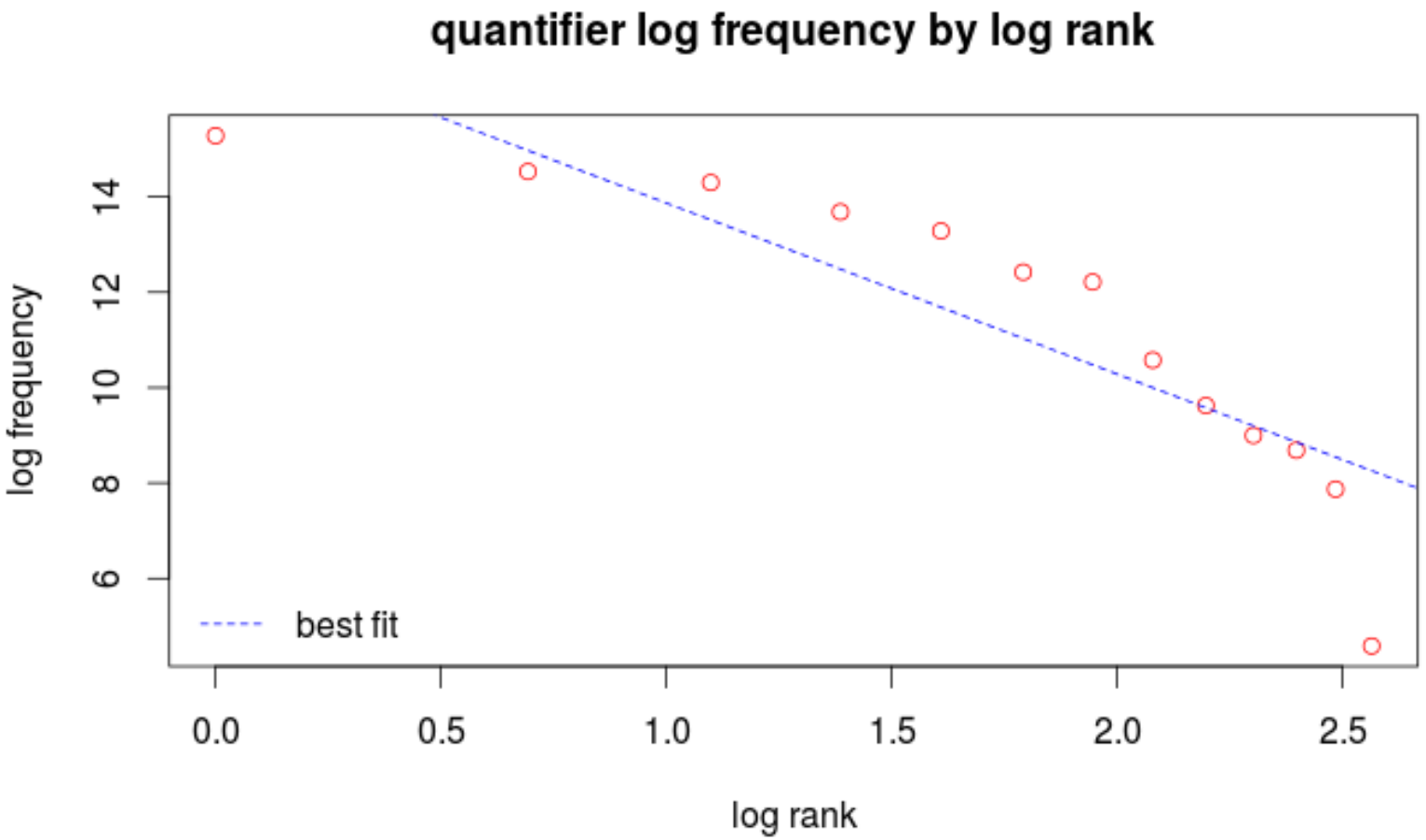
- GQ class not very explanatory overall ($p = 0.135$, AOV)



- GQ complexity not very explanatory overall either ($p = 0.154$, AOV)



- GQ length explanatory overall ($p = 0.00745$, AOV)
- but complexity weakly influences the distribution of GQs of the same length ($p = 0.0596$, AOV)



- log-log (ln-ln) regression (to test for power law):

$$R^2 = 0.7581$$
$$p = 0.0001074$$

Open Problems

Q1: What is exactly a quantifier?

Q2: Is it possible to define a distributional concept of GQs, to observe and count all GQs in a corpus?

Q3: Would (expressive) pattern matching languages yield more accurate results?

Q4: Quantifier expressions are **(1)** very ambiguous and **(2)** non compositional. How can we best analyze annotated corpora given constraints **(1)** and **(2)**?

Q5 Large corpora are (semiautomatically) annotated. How do we deal with annotation noise, especially when looking for log-tail quantifiers?


Thank You!

References I

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta.

The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora.

Language Resources and Evaluation, 43(3):209–226, 2009.

John Barwise and Robin Cooper.

Generalized quantifiers and natural language.

Linguistics and Philosophy, 4(2):159–219, 1980.

Richard Montague.

Universal grammar.

Theoria, 36(3):373–398, 1970.

Richard Montague.

The proper treatment of quantification in ordinary english.

In *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, 1973.

References II

Jakub Szymanik.

Quantifiers in Time and Space.

Institute for Logic, Language and Computation, 2009.

Camilo Thorne and Jakub Szymanik.

Quantifier distribution and semantic complexity.

In *Proceedings of the 2013 Tbilisi Colloquium (TbiLLC 2013)*, 2013.

Camilo Thorne and Jakub Szymanik.

Semantic complexity of quantifiers and their distribution in corpora.

In *Proceedings of the 11th International Conference in Computational Semantics (IWCS 2015)*, 2015.